Simplifying Mainframe Data Access with IBM InfoSphere System z Connector for Hadoop
IBM Redbooks Solution Guide

Did you know?
For many, the IBM z Systems™ mainframe forms the backbone of mission-critical business applications and business processes, securely and reliably storing and processing massive volumes of data, day after day. It is estimated that 80% of the world’s corporate data resides or originates on mainframes. Mainframes provide exceptional qualities of service, such as high availability, security, systems integrity, ability to handle mixed workloads and cope with unexpected workloads. These solutions are now essentially table stakes as businesses advance to the next round and look for new competitive advantages in big data. A Gartner survey noted that 79% of respondents cited transaction data as a data type being used in big data projects (source: Gartner Inc, Research Note G00263798 - Survey Analysis: Big Data Investment Grows but Deployments Remain Scarce in 2014, Nick Heudecker, Lisa Kart, Date Published 09/09 2014). With big data, organizations mine more data (orders of magnitude more) for insights into customers and processes in order to gain new advantages.

Several industry trends are driving big data and affecting how organizations think about data:

- Plummeting cost of storage, now at a commodity level (the cost of 1 TB of storage is in the range of USD 50)
- Relative ease of data capture, and increasing availability of data in electronic form
- Increasing creation and capture of nontraditional data types (such as log files, images, audio, documents, and unstructured text)
- Rapid and accelerating growth in the volumes and variety of data (which is a function of the previous three trends)
- The need to accelerate insight closer to real-time analysis

As a result of these trends, organizations face a complex set of challenges in infrastructure and data architecture. They need mainframe data access strategies that continue to satisfy existing operational challenges, while taking advantage of innovations and emerging opportunities to meet new requirements.

Business value
The need for access to mainframe data is not new. Reporting systems and data access tools have existed for decades. The business value of new approaches to mainframe data access comes as a result of enabling a platform for secure and more cost-effective analysis of data, while simultaneously supporting a broader set of applications in support of new business initiatives.

Much of the discussion about big data has been about the expanding universe of data from social media, new applications, and devices (and schemas). Although these new sources are valuable, organizations with mainframe servers have a unique store of data that can be more fully leveraged by using big data technologies. These technologies can be used to tackle new challenges and solve old ones more efficiently.
• Enable data-driven decisions
  From point-of-sale special offers to strategic decisions about new service offerings, all levels of an organization can benefit from better information. Mainframe-resident data is increasingly needed to support downstream analytic models.

• Self-service access to analytics
  To support planning and decisions, information must be timely. Lengthy business intelligence projects can preclude the iterative queries that are sometimes part of finding an optimal or innovative solution.

• Efficient allocation of skills
  Organizations need to deploy scarce talent efficiently. They are looking for ways to answer needs from the business without the need for specialized expertise or custom programming.

• Speeding the delivery of new applications
  Businesses are under pressure to deliver new applications through new business channels, such as mobile, involving mainframe data while minimizing application maintenance costs.

• New application types
  Business units are increasingly interested in combining mainframe transactional data with data from other sources to improve customer service, gain efficiencies, and support the creation of new service offerings.

As a result, there is a need for a faster feedback loop to enable faster decisions and greater agility. Organizations that can quickly react to changing conditions, respond first to what their customers want, and monitor and make mid-course corrections to processes will have the greatest competitive advantages.

Solution overview

Although no amount of technology can solve every problem, Hadoop gets a lot of attention for new workloads. Although some associate Hadoop with new data types (such as video, text, or social media graphs), the data reality is different. A 2014 Gartner research note found that when asked “Which types of big data does your organization currently analyze?” 79% of respondents cited transaction data, followed by 58% citing log data, both of which are data types common in mainframe environments (Survey Analysis: Big Data Investment Grows but Deployments Remain Scarce in 2014, published 9 September 2014).

A discussion of Hadoop is beyond the scope of this paper, but it is worth reviewing the following list for some of the properties that make it interesting to IT professionals:

• Capacity and scalability. As data sets grow into the petabytes, Hadoop helps take advantage of plummeting storage costs with a unique computing architecture that scales performance and capacity linearly (and affordably) with additional hardware.

• Open standards-based. Hadoop is essentially open at its core, with all vendors building on core Apache Hadoop components. This makes Hadoop a low-risk technology that can be sourced in a competitive environment with little risk of proprietary lock-in to a particular vendor.

• Hadoop is now "good enough" for many use cases. This is perhaps the most important point. In the past, taking on a Hadoop project required deep skills in Java. Now, the tools in Hadoop distributions are much more accessible. Although Hadoop is not going to replace transactional systems, optimized columnar databases, or traditional enterprise data warehouses anytime soon, for many applications it is now up to the task, and the economic benefits are compelling.
As Hadoop continues to mature into an important investigative analytical technology in the modern data center, quality solutions for efficient, high-speed connectivity between the mainframe and Hadoop clusters becomes essential.

The IBM System z Connector for Hadoop

The IBM InfoSphere® System z Connector for Hadoop is software that provides fast and seamless point-and-click data integration between mainframe sources and destinations (including Hadoop clusters and other environments).

Customers can easily extract data from IBM z/OS® sources without the need for mainframe-based SQL queries, custom programming, or specialized skills. After the data (and metadata) is in Hadoop, clients can use tools in the Hadoop platform for processing and analytics. Metadata is preserved, so data can be directly inserted into Hive for access through HiveQL (Hadoop’s SQL subset).

Hadoop processing can take place on an external cluster connected to the IBM zEnterprise® System mainframe, or directly on mainframe Linux partitions using the IBM Integrated Facility for Linux (IFL) for added security. This is a requirement for sensitive and personal identifiable information (PII).

The point-and-click GUI provides access to the following components:

- IBM DB2®
- VSAM
- QSAM
- IBM Information Management System (IMS™)
- IBM z/OS Management Facility (z/OSMF)
- IBM Resource Measurement Facility (RMF™)
- System log files and operator log files

The unique EL-T (extract, load, and then separately transform) architecture provides secure, agile, access to mainframe data at lower cost by eliminating staging and MIPS for data conversion. This offers the following key advantages included:

- Point-and-click integration, with near real-time access to mainframe-resident data
- Self-service data access without a need for mainframe system programming
- No MIPS used for data conversion or SQL extracts (binary extraction is used)
- In-flight data format conversion with no intermediate data staging
- Minimal load on the z/OS environment, so that data transfers do not interfere with other mainframe workloads
- For security, support for IBM Resource Access Control Facility (RACF®) controls to data access and encryption of in-flight data

In later sections, this guide describes some of the capabilities of the System z Connector in more detail.

Solution architecture

The IBM InfoSphere System z Connector consists of a set of services that run mainly on a Linux node, along with a few components installed on z/OS. The architecture is flexible, and the Linux node may be a discrete Intel, an IBM Power® based system, or a virtual machine running Linux on a z Systems IFL.

The System z Connector copies data from a defined source (usually z/OS) to a target (usually Hadoop). On the mainframe, data is extracted in raw binary and streamed to the target platform without creating temporary (staging) copies of the data. This extraction technology is one of the key reasons that the System z Connector is efficient and consumes minimal mainframe resources.
Near wire-speed performance

The usual method (without the System z Connector) of extracting data from DB2 on z/OS is to use a SQL query, which results in considerable mainframe processing. Queries need to be interpreted, processing occurs, data is read, and query results are written to DASD. The System z Connector skips all of these steps. It reads directly from the binary DB2 data source and streams the binary data from the mainframe to a Linux node where the binary data stream is converted in memory. This approach is the key reason that the System z Connector consumes minimal z/OS MIPS and no mainframe DASD.

Figure 1 shows an architectural diagram, and descriptions of the key components follow.

![Diagram showing the process using the System z Connector](image)

**Figure 1. How the process works using the System z Connector**

The three major components are the Management Console (client applet), vHub, and vConnect:

- **Management Console.** The graphical user interface (GUI) that manages the user interactions with the System z Connector for Hadoop system. It is a multi-tenant UI which is based on an applet/servlet interface. The Java applet executing in the browser is served by the Java Platform, Enterprise Edition (JEE) server (Tomcat).

- **vHub.** The component that is at the heart of the System z Connector. It runs on Linux and first uses the vStorm Connect facility to access the source data on z/OS via vConnect agents deployed on the z/OS environment. Second, as the data is streamed to vHub, it converts the data to the target format dealing with issues like EBCDIC to ASCII conversion in various code pages and packed decimal data types. Finally, vHub stores the data in the target big data platform.

- **vConnect.** A term used to describe a set of connectors that work with vHub.

The following data sources are supported by the System z Connector:

- IBM DB2 databases, tables, and columns, including support for predicates to filter rows
- IBM IMS™
- System log (SYSLOG) with filters
- Operator log (OPERLOG) with filters
Simplifying Mainframe Data Access with IBM InfoSphere System z Connector for Hadoop

- System Measurement Facility (SMF), record types 30 and 80
- Resource Management Facility (RMF)
- JDBC (for other relational database management systems, or RDMses)
- VSAM and QSAM (sequential) without custom programming for COBOL copybooks (supports data item subsets and multiple copybooks)

vConnect can transfer data by using SSL encryption. Transfers to Hadoop running on Linux outside of z Systems can use multiple Gigabit Ethernet interface,s and transfers to Hadoop on Linux on z Systems can use the IBM HiperSockets™ interface for very high-speed data movement.

On the other side of the vHub, the following components are the targets to which the System z Connector can move data:

- **HDFS.** The Hadoop Distributed File System. The System z Connector will interface to the HDFS NameNode and move mainframe data as comma-separated values (CSVs) or Avro files directly into to the Hadoop Distributed File System. Avro files are compressed to reduce storage requirements.

- **Hive.** Metadata is written to the Hive server, reflecting the source data schema, and file data is moved to HDFS to make data available on Hadoop for HiveQL queries. Hive is not a “data format” in the sense that a relational database has its own on disk storage format. Instead, Hive is a facility that enables a schema to be imposed on existing data in HDFS so it can be queried by using HiveQL, an SQL-like language. Data that exists in Hive is already stored in HDFS. IBM InfoSphere BigInsights™ users can access this data directly using Big SQL, an ANSI-compliant SQL implementation that can directly read and process data stored in Hive.

- **Linux file system.** In addition to landing data in HDFS, the System z Connector can transfer data directly into a Linux file system. The file system can be within the mainframe in the Linux environment or on nodes external to the mainframe. This provides additional flexibility. Data written to a local file system can be used by downstream ETL tools or applications as well as analytics tools and applications. This flexibility is important, because clients might want to move data not only to Hadoop but to other environments also.

- **Network end point.** The System z Connector can also send data to a network end point that is listening on a specific port. The data is made available to the listener as streaming bytes in CSV format. Organizations can build their own applications to parse and process data originating from the mainframe dynamically. If the receiving software can open and read data from a TCP/IP socket connection, it can receive data streamed by the System z Connector.

**Capabilities**

Although not all of the capabilities of the System z Connector are described here, some of the highlights are provided to give readers an introduction to the connector and how it is used.

**Specifying data sources**

There are two steps to specify the data to be transferred. The first is a one-time step to define a new connection to a data source. Data sources for a transfer include data on DASD, UNIX System Services (USS), IBM DB2 on z/OS, or various log files.

Online help is provided through the browser interface so that authorized users can configure data sources for mainframe transfer. Figure 2 shows an example of one of the steps.
Users of System z Connector for Hadoop need to consult with mainframe administrators initially to understand and gather mainframe credentials. After these credentials are provided, users can become self-sufficient in setting up data transfers.

**Specifying data targets**

An interactive wizard is also provided for selecting transfer targets. Targets for mainframe data may be the HDFS on Hadoop clusters, Hive tables, a Linux file system, or a network end point where a service is listening on a TCP/IP port.

With the System z Connector for Hadoop, there can be multiple concurrent connection definitions. Therefore, it is possible to move data in the zEnterprise Systems environment not to just one target, but multiple targets, including multiple Hadoop clusters.

For example, a user may elect to move sensitive data to a Hadoop cluster configured on Linux partitions within the z Systems environment (on IFLs) and other less sensitive data to an external Hadoop cluster.

**Filtering transferred data**

When transferring data from the mainframe, you are often interested only in a subset of the rows or columns from a given data source. Rather than transfer the whole data set and filter it on the receiving Hadoop cluster, the System z Connector allows filtering of data dynamically.

As Figure 3 shows, data transfers can be configured to select individual data columns to be transferred or to filter rows based on certain criteria. This improves flexibility by ensuring that you are transferring only the required data.
Automated scheduling of transfers

Data transfer can be interactive or scheduled for automated transfer. For example, mainframe logs or customer transaction data could be moved daily to the Hadoop cluster for downstream processing. System z Connector for Hadoop allows users to configure automated transfers that repeat. By using automatic transfers, workloads are reduced, as is the opportunity for human error that often occurs with manual processes. Furthermore, the GUI may be used to define and test a particular type of data transfer. After it is defined, transfers can be invoked outside of the GUI. Transfers can be scripted or can run under the control of a mainframe job scheduling system. This is an important capability for sites with complex requirements that might need to transfer thousands of files daily. The capabilities to configure data sources and targets and to specify how the data is configured help make it easy to transfer data from various mainframe sources to multiple destinations.

Use scenarios

To illustrate a scenario where it may be necessary to combine mainframe data in Hadoop with data from other sources, this guide focus on a hypothetical retailer. The example is applicable to a broad range of industries and similar use cases are found in telecommunications, healthcare, insurance, banking, and other industries.

Understanding the business need: An international retailer

The retail business is complex, particularly in the last decade with increasingly connected consumers, the mobile phenomenon, intense competition, and fast-moving product cycles and trends.

Major retailers frequently operate in many countries and distribute products from a broad set of suppliers and manufacturers through multiple sales channels. The sales channels include retail locations, catalog stores, and country-specific e-commerce websites. For many retailers, key operational data exists on a
corporate mainframe. This includes operational and transactional data related to customers, suppliers, stock items, inventory levels, and more.

Among the many challenges that retailers face is predicting and maintaining optimal levels of inventory across warehouses and distribution centers in various geographies. If a retailer has inadequate supply before seasonal spikes, such as Black Friday or Christmas shopping season, there can be a significant lost opportunity cost. Just as problematic, retailers with excess inventory face increased carrying costs and potential restocking charges if goods are not moved quickly.

Mainframe systems have long been augmented by data warehouses and decision-support systems (DSS) to help address these challenges. Data warehouse environments are frequently subject to many complex queries. Organizations that can operate from a larger set of data and who have good predictive models have a significant advantage over competitors.

An example of a query against a data warehouse might be to assess the degree to which a change in a pricing strategy will affect inventory for various items (this specific example is drawn from the Transaction Processing Council's TPC-DS benchmark at http://tpc.org). The 99 business-oriented queries in the TPC-DS benchmark are designed to be representative of the broad sets of queries that major retailers typically run, such as:

For all items whose price was changed on a given date, compute the percentage change in inventory between the 30-day period BEFORE the price change and the 30-day period AFTER the change. Group this information by warehouse.

Other queries of interest to consider are item color and size preferences by demographic and geography, with various constraints or items that are most frequently returned, sorted by sales outlet, product class, or manufacturer to estimate the effect of stocking these frequently returned items on profit margins.

Managing inventory has become more challenging than ever

Although a data warehouse is essential to making good decisions, a challenge for most organizations is that the analysis previously described is based on history. Decision-support systems and business analytics engines can query and analyze only data that is in the data warehouse.

In the age of big data, these queries might be missing “forward-looking” data points and nontraditional data sources. If those are incorporated into the analysis, the information can help forecast required inventory and predict what might happen next with more certainty and precision.

Consider the following examples:

- Web browsing behaviors and search terms parsed from server logs can provide insight into customer interest and help predict demand for specific product categories in specific geographies.
- Analysis of customer service channels (such as chat, email support, and recorded call center conversations) can provide early warnings about products or manufacturers with quality problems that are likely to lead to future returns.
- Analyzing publicly available data from social media can reveal deep insights about customer sentiment and trending topics.
- Using web-crawling techniques can help retailers understand what competitors are offering in geographies where they do business, which helps them price products more effectively and predict sales volume.

Although this guide has focused on examples related to inventory levels, this is really just the tip of the iceberg. The real opportunity that arises from having a more complete view of your customer is to provide tailored offers to customers or communities of customers at the time they are most receptive to the offers.
IBM clients might be interested in using social media data to assess the effectiveness of advertising campaigns, or they might use dynamic pricing algorithms based on customers’ patterns of access to a website or their purchase histories. As an illustration, Figure 4 shows a graphic that represents tweets about a particular product to help a retailer understand what products are of interest to prospective customers and where.

Figure 4. World map and graph that show tweets by months

**A better approach is required**

The previous example illustrates the need to integrate information from existing and new, nontraditional sources. Because of its ability to handle large and diverse data types cost-effectively, Hadoop is often where the following diverse data sources converge:

- Customer demographic and transactional data from the mainframe
- Historical data from operational warehouses
- Supplier and store-level data from various sources
- Social media data gathered from Twitter and public data aggregators
- Web crawlers that track competitive promotions and prices
● Email and chat data from customer service centers
● Recorded call center conversations converted to text
● Consolidated log files from the various services that comprise the e-commerce infrastructure that supports various locations
● Geographic data from mobile devices that might indicate local trends

Streamlined data transfer from the mainframe has become essential

To support the new types of applications described previously, timely access to information from all data sources, including the mainframe, is essential. As Hadoop-based tools are increasingly used to view and analyze data in new and interesting ways, there is more pressure to quickly provide continuous access to data in required formats. It is no longer practical to address requirements individually and engage a system programmer to help with every request.

Mainframe data transfers to Hadoop need to be able to address the following concepts:

● **Self-service.** To not require the engagement of people with mainframe expertise
● **Secure.** To ensure that truly sensitive data is not inadvertently exposed in less secure environments, such as Hadoop
● **Automated.** To run automated data transfers, without operator intervention (to do so, downstream models will require continuous access to the latest data)
● **Flexible.** For transfers to draw data from multiple z/OS data sources and send data to multiple systems, both Hadoop and non-Hadoop
● **Efficient.** For data transfers to be fast and efficient and avoid the need for intermediate data storage and processing to keeps costs down and avoid unnecessary copies of data

Other common use cases

Two other common use cases for the System z Connector for Hadoop are described in the following subsections:

● **ETL processing offload**

ETL processing is a common use of Hadoop. Offerings such as IBM InfoSphere® DataStage® are well-suited for ETL requirements that involve structured data on the mainframe, Hadoop might offer better ETL for other data types. Often, ETL is described as “ELT” in Hadoop environments, reflecting the fact that transformation operations are performed after data is loaded into Hadoop.

Performing transformation operations in Hadoop has many distinct advantages, including self-service access to data, inexpensive archiving, and the following benefits:

○ **Reduced processing times and costs**
  Very large data sets benefit from the parallel processing facilities inherent in Hadoop and the corresponding reduction in costs.

○ **Mainframe offload**
  The lower cost of processing and storage in Hadoop is even more apparent in cases where batch processing is offloaded from the mainframe. This frees up mainframe MIPS for mission-critical workloads.

○ **Innovative ETL tools**
  The Hadoop ecosystem is evolving rapidly, bringing with it new applications for ETL operations and an increasing range of analytics applications with transformation functions built in.

○ **Greater access to talent pools**
  Hadoop is built upon technologies that have been broadly adopted, including Linux, Java, C++, and XML.
Support for a broad variety of data
Many data sets can be accumulated in Hadoop (this is sometimes referred to as a *data lake* or *data hub*), even if each set has a different structure or no structure at all. This can include nontraditional binary formats, such as images, sounds, and videos, and unstructured texts, such as PDF files and email.

**Mainframe log file analysis**

Another important use for Hadoop and the mainframe is the analysis of various mainframe log file formats. IBM z/OS Management Facility (z/OSMF) provides a standardized method for writing activity to a data set. z/OSMF provides complete instrumentation of baseline activities on the mainframe, including I/O, network activity, software use, processor use, and more. Add-on components, including IBM DB2, IBM CICS®, IBM MQ, and IBM WebSphere® Application Server provide their own log file-type reporting by using z/OSMF.

Hadoop, and BigInsights in particular, provide rich facilities for parsing, analyzing, and reporting on log files of all types. When clients analyze logs by using tools in BigInsights, they can gain several benefits:

- Understand user, application, and group usage patterns
- Identify issues before they affect production applications
- Gather trend information that is useful for capacity planning
- Find intrusion attempts, other security issues, or evidence of fraudulent activity

Because Hadoop is designed to support large data sets, clients can retain raw log data for longer than might be feasible otherwise. More data helps clients discover longer-term trends that are related to use and variance in activities.

**Integration**

**Deployment models**

The System z Connector for Hadoop supports several different deployment models. The appropriate model will depend on the client’s environment and can be affected by several considerations.

The source for the connector is generally z/OS based data (although the connector can be used to source data from other sources including JDBC). There are several options however related to target environments, including those noted in the following list:

- InfoSphere BigInsights (Hadoop) running on the mainframe
- InfoSphere BigInsights running on local infrastructure
- InfoSphere BigInsights running on IBM or third-party cloud services
- Third-party Hadoop distributions (such as Cloudera or Hortonworks) running on local infrastructure
- Third-party Hadoop distributions running on IBM or other cloud services

This guide considers the benefits and limitations of these different approaches briefly in the following sections.

**InfoSphere BigInsights on the mainframe**

IBM InfoSphere BigInsights is the IBM enterprise-grade Hadoop offering. It is a complete Hadoop distribution that provides the same standard Hadoop components available in other Hadoop distributions and with additional features. It can run on standard Intel hardware and on Linux for z Systems’ guests running in the virtualized mainframe environment. In this environment, each virtual machine that is
configured in the IBM z/VM environment corresponds to a Hadoop node. The System z Connector can connect directly to the BigInsights cluster that is running in the Linux for z Systems environment and use HiperSockets for a high-performance, secure connection between the z/OS and Hadoop environments.

Mainframe managers find this approach attractive in the following situations:

- They are dealing with sensitive data and want to keep all processing within the security perimeter of the mainframe
- Most of the data being processed originates on the mainframe
- Data sets are large but small enough to be handled economically on the mainframe (tens of terabytes as opposed to petabytes)
- They want to take advantage of integration features between InfoSphere BigInsights and mainframe tools, such as DB2

**InfoSphere BigInsights on a separate local cluster**

IBM InfoSphere BigInsights can also be deployed on commodity Intel-based clusters and connected to the mainframe by using one or more 10 GbE connections. This approach can be advantageous to clients who are in the following situations:

- They have the in-house capacity to manage a distributed cluster environment discrete from the mainframe
- They are comfortable with moving copies of mainframe data off the mainframe onto a local cluster
- Most of the data volumes are originating from the mainframe
- The environment is expected to grow very large (hundreds or terabytes or petabytes), and they want to take advantage of commodity components

**InfoSphere BigInsights in the cloud**

In addition to being deployable on premises, IBM InfoSphere BigInsights can be deployed in the cloud. *IBM BigInsights on Cloud* is a service that enables IBM clients to get started quickly on a high-performance bare-metal infrastructure while avoiding the cost and complexity of managing Hadoop clusters on their own premises.

As a part of the IBM relationship with Twitter, select configurations of IBM’s Cloud Services include Twitter’s Decahose service, along with an application that makes it easy to incorporate up-to-date Twitter data into analytic applications on the Hadoop cluster.

Clients find this approach attractive in the following situations:

- Much of the data originates in the cloud or originates outside of the organization (for example: social data, data from external aggregation services, or data feeds).
- The client does not want to manage local infrastructure.
- The client wants to have a variable-cost model, where they can adjust capacity (up or down) rapidly as business requirements change.
- The client is comfortable moving corporate data up to the dedicated infrastructure on the cloud service, or their analytic requirements are such that they can avoid the need to do this.

**Third-party Hadoop distributions on premises**

Many IBM clients have made corporate decisions based on Hadoop and already have Hadoop clusters deployed on their premises. The System z Connector for Hadoop uses standard Hadoop interfaces, so from a technical standpoint, it should be straightforward to connect to open source or commercial Hadoop clusters. Popular third-party Hadoop environments, including Cloudera and Hortonworks, are supported by IBM. (It is important to check that the specific third-party Hadoop environment is supported, because support can vary depending on the version of the Hadoop distribution.)
This approach is attractive when you are in the following situations:

- You have already standardized on a third-party Hadoop environment.
- You do not see a need to run Hadoop components on the mainframe.
- You do not require the value-added capabilities in IBM’s BigInsights Hadoop distribution.

**Third-party Hadoop distributions in the cloud**

Just as there are a variety of Hadoop solutions that can be deployed on premises, there is even a wider variety of cloud providers who offer Hadoop services. The decision about what provider to select is complex. There are many factors to consider, but assuming that the third-party provider exposes standard Hadoop services, the IBM InfoSphere System z Connector should work. It is important to check with IBM to be sure that the System z Connector is supported with the chosen third-party Hadoop-as-a-service offering. This approach is appropriate when dealing with the following situations:

- You have already selected a preferred Hadoop-in-the-cloud provider.
- You are comfortable with moving mainframe data to an external cloud provider, or the application requirements are such that there is not a need to do so.
- The data sets originated primarily in the cloud or the data set sizes are small enough to make network transfer from the mainframe provider to the cloud provider practical.

**Hybrid environments**

In the foreseeable future, there might not be a single data hub or data lake. Real data center environments are often complex, supporting multiple applications and user communities, and hybrid environments can emerge as the most practical solution. Organizations may elect to have separate Hadoop clusters deployed on the mainframe, on local infrastructure, and with one or more external cloud service providers. Clusters may support different applications or different lines of business with different business and technical requirements.

Consider the following example:

- A business has most of their customer and transaction data on the mainframe. The data is sensitive and includes financial information. They want to maximize Hadoop-based tools to parse, transform, and cleanse the data to remove identifiable information before they are comfortable moving the data outside of the mainframe security perimeter.
- They are also deploying a new sentiment analysis application that will incorporate Twitter data. Because the Twitter data is from the cloud, it will be faster and more cost-effective to process the raw Twitter data on a cloud-based cluster and move summary data back to on-premise systems.
- The client also has additional data types, including web-server logs, chat logs, and email data, that they would like to analyze together with summary social data and cleansed data coming from the mainframe. To achieve this requirement and to archive the data over time, it makes sense for the client to deploy a third cluster to combine and retain the summarized data coming from the two other clusters (such as the mainframe cluster used for processing and cleansing customer transactions and the cloud-based cluster processing social media data).

For such an example, it is perfectly reasonable to have three separate clusters that support different business requirements. One attraction of using BigInsights as a standard Hadoop deployment is that clients can use the software across all three tiers to simplify the environment and improve administrative efficiency.

As mentioned previously, analysis of social data can be one of the applications best suited to deployment in the public cloud because the data is created in the cloud, outside of the corporate firewall, and there are fewer privacy concerns. With this in mind, IBM is developing solutions that make it easier than ever for clients to start incorporating social data into their analytic applications.
Specific configurations of IBM BigInsights on Cloud (the IBM enterprise Hadoop-as-a-service offering) feature a built-in service that interfaces with a provided Twitter Decahose service. This provides clients with immediate access to valuable Twitter data and has tools to store and analyze the data. Although social media data from various sources including Twitter can be accessed and analyzed on any major Hadoop distribution by including Twitter data as a part of the service, IBM helps make it easier to get productive quickly.

**Supported platforms**

The IBM InfoSphere System z Connector for Hadoop runs on Linux hosts (on or off the z Systems environment).

Supported Hadoop environments that can be targets for the System z Connector for Hadoop include those in the following list:

- IBM InfoSphere BigInsights for Linux on z Systems (version 2.1.2)
- IBM InfoSphere BigInsights on Intel Distributed Clusters
- IBM InfoSphere BigInsights on Power Systems
- IBM BigInsights on Cloud
- On-premises Cloudera CDH clusters
- On-premises Hortonworks HDP clusters
- Apache Hadoop for Linux on z Systems
- Veristorm zDoop (open source Hadoop offering for z Systems Linux)
- Veristorm Data Hub for Power Systems (Hadoop distribution for Power Systems)

The System z Connector components on the mainframe will normally be deployed on Integrated Facility for Linux (IFL) processors on z Systems environments.

Information about how Linux is deployed on IBM z/VM environments can be found in the IBM Redbooks publication titled *The Virtualization Cookbook for IBM z/VM 6.3, RHEL 6.4, and SLES 11 SP3*, SG24-8147: [http://www.redbooks.ibm.com/abstracts/sg248147.html](http://www.redbooks.ibm.com/abstracts/sg248147.html)

**Ordering information**

The System z Connector for Hadoop is licensed on a per virtual machine basis.

Two different editions of the connector are available:

- **The IBM InfoSphere System z Connector for Hadoop BigInsights Edition**: Provides the capability to target IBM InfoSphere BigInsights clusters. These can be distributed clusters deployed on Intel or IBM Power Systems or can be cloud-based clusters configured as part of the IBM BigInsights on Cloud service.

- **The IBM InfoSphere System z Connector for Hadoop Enterprise Edition**: Entitles users to transfer data to third-party Hadoop clusters, including Cloudera or Hortonworks. These clusters are generally deployed on Distributed Linux systems. They can also be deployed as a service on third-party public clouds. Check with IBM about the support status of specific third-party cloud providers.

Table 1 provides essential ordering information.

<table>
<thead>
<tr>
<th>Description</th>
<th>Part Number</th>
</tr>
</thead>
</table>

Simplifying Mainframe Data Access with IBM InfoSphere System z Connector for Hadoop
**IBM Passport Advantage®**

<table>
<thead>
<tr>
<th>Product Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM InfoSphere System z Connector for Hadoop BigInsights Edition Virtual Server Subscription and Support 12 Months</td>
<td>D1B26LL</td>
</tr>
<tr>
<td>IBM InfoSphere System z Connector for Hadoop BigInsights Edition Virtual Server Annual SW Subscription and Support Renewal</td>
<td>E0KIHLL</td>
</tr>
<tr>
<td>IBM InfoSphere System z Connector for Hadoop BigInsights Edition Virtual Server SW Subscription and Support Reinstatement 12 Months</td>
<td>D1B2CLL</td>
</tr>
<tr>
<td>IBM InfoSphere System z Connector for Hadoop Enterprise Edition Virtual Server License + SW Subscription and Support 12 Months</td>
<td>D1B2DLL</td>
</tr>
<tr>
<td>IBM InfoSphere System z Connector for Hadoop Enterprise Edition Virtual Server Annual SW Subscription and Support Renewal</td>
<td>E0KIILL</td>
</tr>
<tr>
<td>IBM InfoSphere System z Connector for Hadoop Enterprise Edition Virtual Server SW Subscription and Support Reinstatement 12 Months</td>
<td>D1B2ELL</td>
</tr>
</tbody>
</table>

**Media packs**

<table>
<thead>
<tr>
<th>Product Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM InfoSphere System z Connector for Hadoop BigInsights Edition Version 1.1.0 Multiplatform English Media Pack</td>
<td>BB1J0EN</td>
</tr>
<tr>
<td>IBM InfoSphere System z Connector for Hadoop Enterprise Edition Version 1.1.0 Multiplatform English Media Pack</td>
<td>BB1J1EN</td>
</tr>
</tbody>
</table>

**Charge metric**

<table>
<thead>
<tr>
<th>Product Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM InfoSphere System z Connector for Hadoop - license per virtual server</td>
<td>5725-S33</td>
</tr>
</tbody>
</table>


**Related information**

For more information, see the following resources:

- IBM InfoSphere System z Connector for Hadoop product page [http://ibm.co/1CpH79P](http://ibm.co/1CpH79P)
- IBM InfoSphere System z Connector for Hadoop enables mainframe users to harness the power and cost-efficiencies of Hadoop with IBM z/OS data (IBM United States Software Announcement 214-343)
http://ibm.co/17eTGfR

- IBM Redbooks publications:
  - *Hadoop and System z*, REDP-5142

- IBM Offering Information page (announcement letters and sales manuals):

  On this page, enter product name, select the information type, and then click **Search**. On the next page, narrow your search results by geography and language.
Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service. IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you. This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk. IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

© Copyright International Business Machines Corporation 2015. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.
This document was created or updated on February 20, 2015.

Send us your comments in one of the following ways:
- Use the online Contact us review form found at: ibm.com/redbooks
- Send your comments in an e-mail to: redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

This document is available online at http://www.ibm.com/redbooks/abstracts/tips1235.html.

**Trademarks**

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml.

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

<table>
<thead>
<tr>
<th>BigInsights™</th>
<th>IMS™</th>
<th>RMF™</th>
</tr>
</thead>
<tbody>
<tr>
<td>CICS®</td>
<td>InfoSphere®</td>
<td>System z®</td>
</tr>
<tr>
<td>DataStage®</td>
<td>Passport Advantage®</td>
<td>WebSphere®</td>
</tr>
<tr>
<td>DB2®</td>
<td>Power Systems™</td>
<td>z/OS®</td>
</tr>
<tr>
<td>HiperSockets™</td>
<td>RACF®</td>
<td>z/VM®</td>
</tr>
<tr>
<td>IBM®</td>
<td>Redbooks®</td>
<td>zEnterprise®</td>
</tr>
<tr>
<td>IBM z Systems™</td>
<td>Redbooks (logo)®</td>
<td>z Systems™</td>
</tr>
</tbody>
</table>

The following terms are trademarks of other companies:

Inc., and Inc. device are trademarks or registered trademarks of Kenexa, an IBM Company.

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.