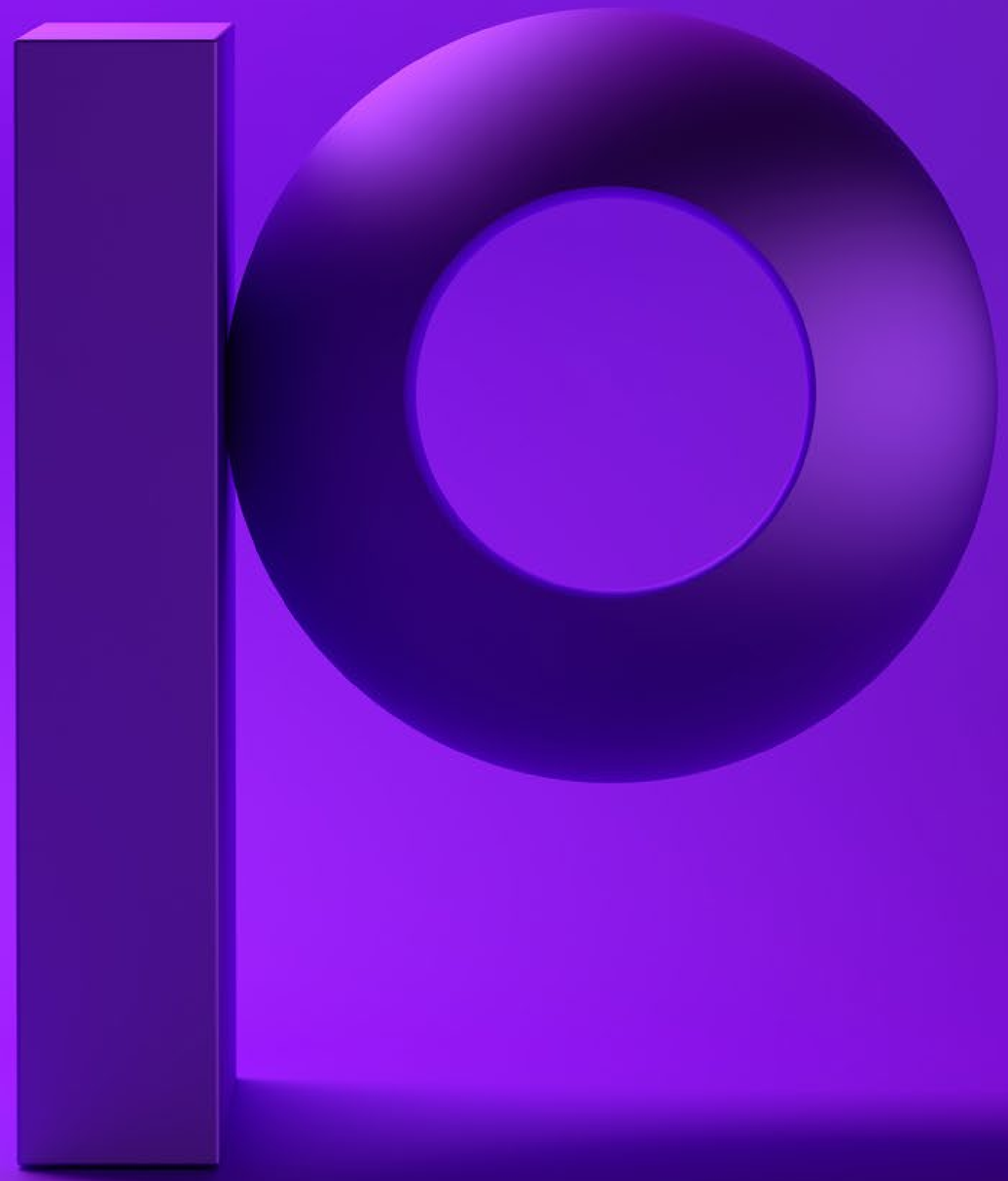
precisely

Mainframe Data for Modern Data Environments:

Best Practices for Bridging the Gap



Mainframes: The Original Big Data

It's been said that 2.5 quintillion bytes of data are generated every day and that number is only predicted to rise moving forward. In a data-rich era dominated by smartphones, Instagram and Facebook, mainframes rarely enter the technology conversation. It's easy to forget that many of the world's largest businesses – financial, retail, healthcare and insurance companies – still generate most of their data from the mainframe. For these companies, and many others, the mainframe is a critical source of big data that cannot be ignored.

Mainframes are the mothership of big data and store about 70% of total structured data in the world. They are the source of priceless transactional data, such as ATM withdrawals, so it would make sense that any meaningful initiative to analyze big data should incorporate mainframe information. The wealth of transactional and other types of data that live on the mainframe is simply too important to exclude.

This eBook will guide you through the process of overcoming the four biggest challenges of leveraging mainframe data, and provide tips and best practices for bridging the gap between mainframes and modern data environments to unlock the value of all your enterprise data.

Big Data and the Big Disconnect

The mainframe delivers extreme performance and scalability, and that's why it commands such a high price premium. The price of the mainframe extends far beyond just the hardware purchase – and with the operation, comes the expenses of processing and storage.

Let's break it down...

Expensive Processing

Unlike any other processing platform, the mainframe is licensed and incurs a fee based on CPU utilization. The more processing you do on the system, the more you pay. It's based on something called MIPS: Millions of Instructions Per Second, a measurement of processing power and CPU consumption.

Operating a mainframe environment could likely utilize 10% to 40% of an organization's IT budget, depending on the size of the mainframe system, and the size of the budget.

Expensive Storage

Mainframes use high-capacity storage to store large files that can be accessed extremely fast, but this high performance doesn't come cheap. Mainframe data is also backed up to tape in case of failure. Because storage costs are so high, only the most recent and relevant data can be stored on the mainframe. Older data is often only stored on tape, freeing up storage for the most recent data on the expensive mainframe, but making older data very difficult and time-consuming to access when it is needed.

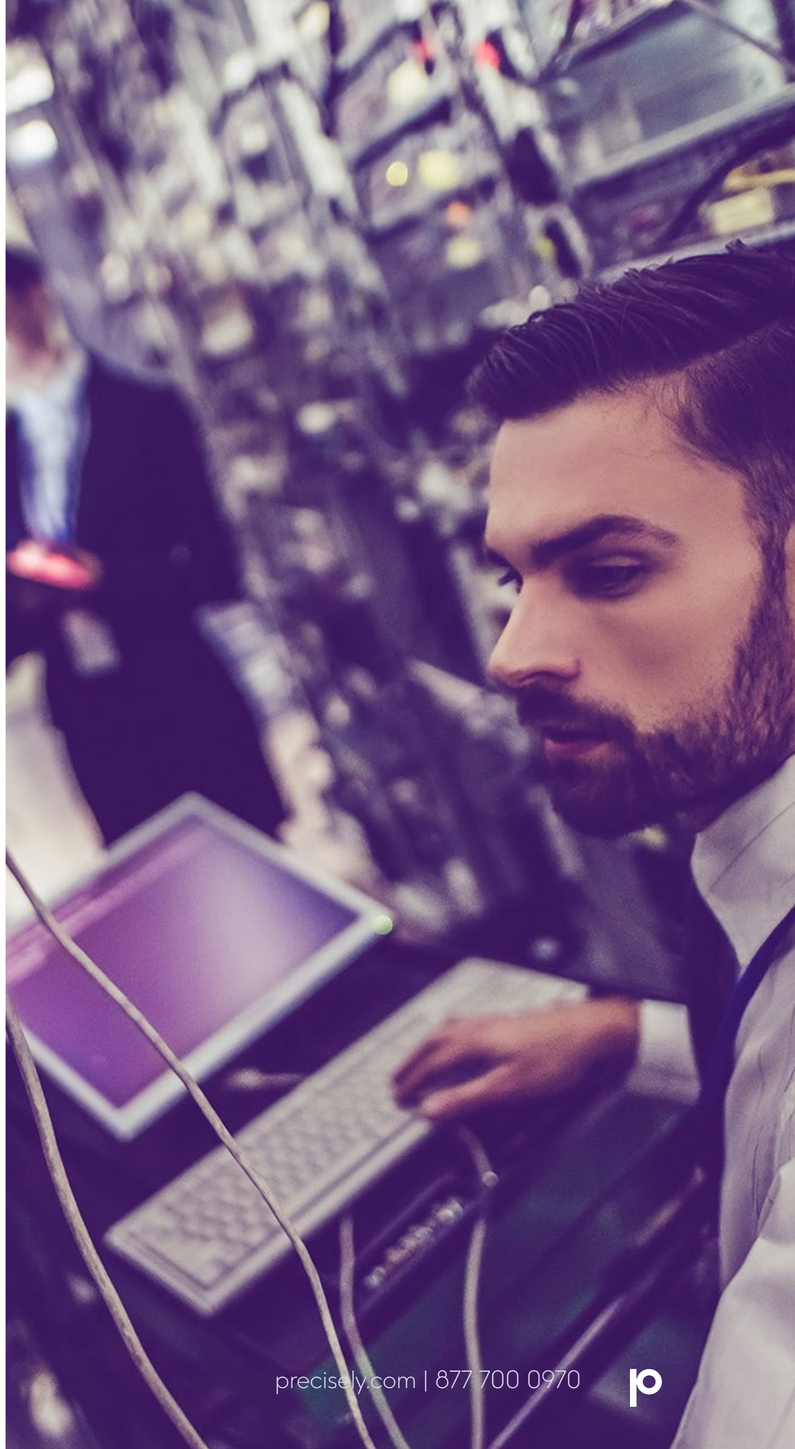
How Can Modern Data Environments Help?

Processing and storage costs are two of the biggest reasons why mainframe data is not used for data intensive work such as long-term trend analysis, or customer lifetime value calculations. This is not just a problem with mainframes. Most traditional data processing architectures carry high storage costs or simply cannot handle large-scale, data intensive analyses.

Enterprises can address both of these challenges by offloading mainframe data and select batch workloads— such as sort, filter, copy and more— to modern data environments that provide highly-scalable processing but with greater flexibility and lower costs. Big data platforms, such as Hadoop MapReduce and Apache Spark, became attractive alternatives to mainframe processing for these reasons. At the same time, cloud computing platforms matured greatly to not only offer flexibility and elasticity, but also enterprise-grade security and a host of applications purpose built for this new environment. Today, many enterprises enjoy the best of both worlds by deploying big data frameworks in cloud and hybrid cloud environments.

By keeping "prime" data storage and processing on the mainframe, and moving less essential data to big data platforms, organizations can make better use of valuable mainframe capacity while improving business service level agreements (SLAs) and significantly reducing costs.

These new technologies allow you to combine all your company's data to gain insight that was previously out of reach, integrating detailed transactional data from mainframes with clickstream data, web logs and social media sentiment data.



Step One: Solving the Integration Gap

While cloud and big data platforms offer many advantages, they don't offer native support for mainframe data.

Perception vs. Reality

Unfortunately, when it comes to moving mainframe data into these modern platforms, the perception of ease is far from the truth. Every file that you want to bring over from the mainframe requires the following steps:

1. Identify and select the data that will be moving. This is something that is not immediately apparent to cloud or Hadoop/Spark developers, most of whom have grown up in a GUI world. Mainframe data may reside in multiple places like VSAM files, or DB2 or IMS databases, and be in the form of flat files, fixed length files or variable length data. So, you should know how to get to the data.

2. Prepare the data. This entails two steps of its own:
    a. Filter the data to get the information you need.
    b. Perform some pre-processing such as joins, sorts and aggregations.

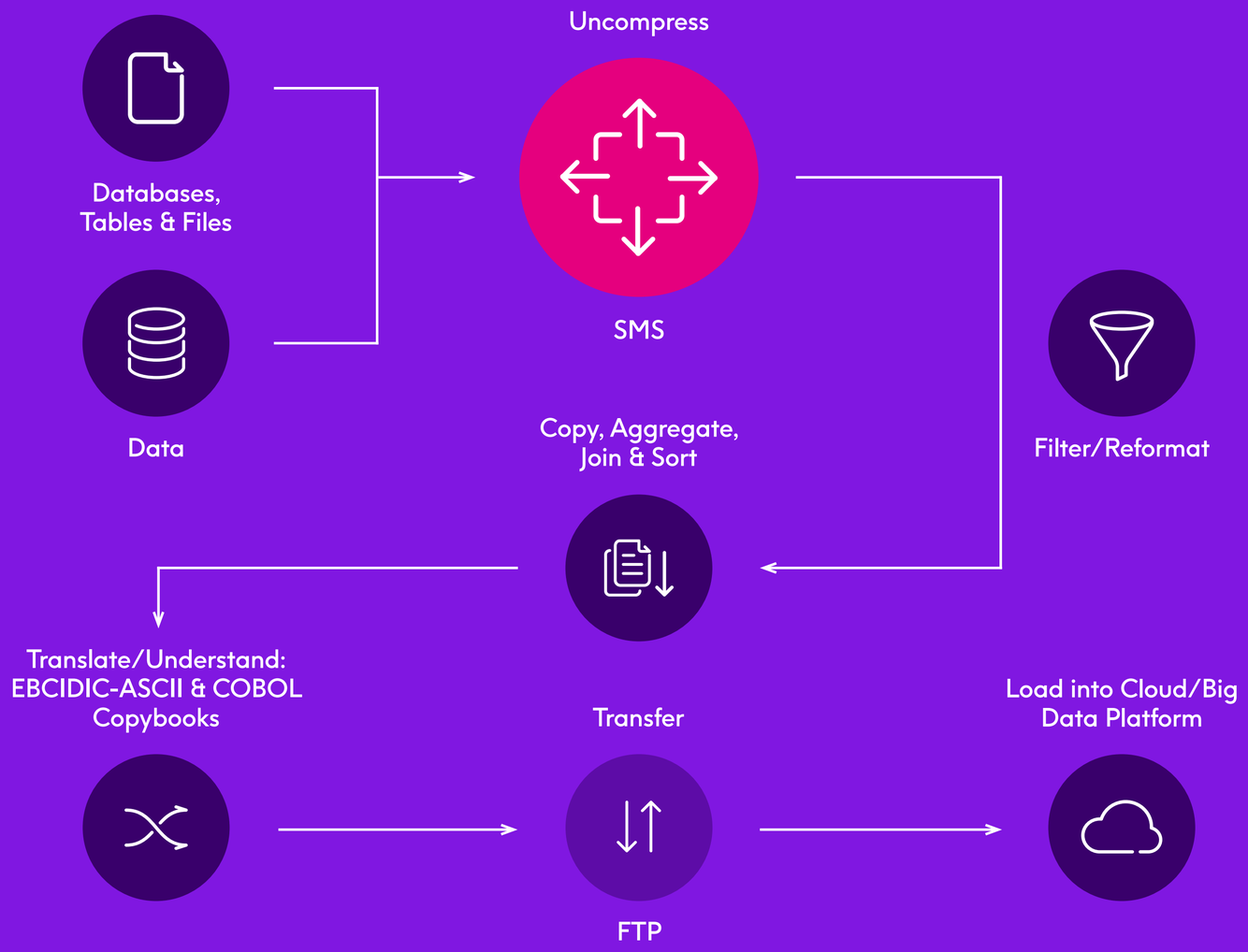
    There is a possibility that the data may be stored in compressed form to save space. Uncompressing the data on the mainframe can result in more CPU time, and MIPS cost money

3. To truly understand the data, you need to understand the data structure– the metadata. On the mainframe, the metadata is stored in dauntingly arcane files called COBOL copybooks, which are complex road maps to the underlying data. Copybooks are always stored separately from the data. This means that over time, the data and metadata can get out of sync, so they don't quite match– and this assumes that one can find the copybooks.

4. Translate the data. From the beginning, mainframe storage was very limited and expensive, so formats like packed decimal, were created to allow more data to be stored in less space.

5. After the data has been rationalized with the COBOL copybook, you can finally perform the FTP transfer into your data lake or enterprise data cloud.

Gap The Long and Winding Road



Mind the Pitfalls Along the Way...

To date, exporting mainframe data and importing it into modern data platforms hasn't just been complicated, but it's been a huge hassle that has had extreme time and cost implications. These have discouraged organizations from using mainframe data for big data and other analytic purposes.

Some pitfalls include:

- CPU time consumption (more MIPS)
- Data management and control issues
- Scaling limitations to meet batch processing windows
- Interfering with regular mainframe production work

Best Practices to Close the Integration Gap

Any solution used to accelerate the access, translation and transfer of mainframe data into modern data platforms, should allow you to:

Connect

- Read files directly from the mainframe including record formats such as fixed, variable with block descriptor and VSAM
- Translate mainframe data in transit, not require data to be staged
- Not install any software on the mainframe

Translate

- Perform data transformation such as, packed decimal, EBCDIC/ASCII, multiple record types, Occurs Depending On arrays, including nested ODO's and more - without coding
- Apply COBOL copybooks to give structure to the data files

Load and Integrate

- Load to distributed file systems and databases
- Load data in parallel, to meet tightening SLAs
- Process data on the cluster or in the cloud, replacing mainframe batch data processing, without writing code
- Perform complex data transformations without coding
- Track end-to-end data lineage for governance and compliance
- Keep data secure by encrypting it both at rest and in flight
- Keep cloud data stores and Hadoop clusters in sync with changes made on the mainframe

Step Two: Bridging the Expertise Gap

Mainframe, cloud and Hadoop/Spark development skills are in such high demand because they are so hard to find – but it's even harder to find developers who understand a combination of these technologies. Just to put things into perspective, COBOL programming appeared in 1959, Hadoop came about in 2006, and cloud computing took off in 2012— that's a gap of about 50 years!

To bridge the gap, it's important that the tools and solutions used to access data, move it into a new platform and develop transformation processes that mimic mainframe batch jobs, can be easily used with minimum training by existing staff. Whether the data is being moved into Hadoop for integration with other data and analysis, or for cost-efficient batch processing, a graphical user interface (GUI) for data flow development is a big plus. In most IT organizations, programming resource time, whether they be COBOL programmers or Java and Scala programmers, will be highly constrained on both teams.

What is a COBOL Copybook?

In the mainframe world, a copybook is a section of code written in COBOL that can be copied from a master and inserted into several different programs, or multiple places in a single program. It is often used to define the physical layout of program data and pieces of procedural code and prototypes.

Copybooks help to:

- Make sure everyone uses the same version of a data layout definition or procedural code
- Make it easier to cross-reference where components are used in a system
- Make it easier to change programs when needed
- Save the programmer time by reducing the need to repeatedly code extensive data layouts

Best Practices to Close the Expertise Gap

Simply ingesting mainframe data into a cloud or Hadoop repository with only open source tools involves a lot of manual effort and coding, but offloading data and batch processing requires organizations to acquire a completely new set of advanced programming skills that are expensive and difficult to find. It's critical to choose a data integration tool that both complements Hadoop and leverages the skills your organization already has, rather than requiring you to hire a new team. When looking for a data integration solution, your options should include:

- A GUI that allows developers to access, translate, load and integrate mainframe data without intervention from mainframe programmers

- The ability to define workflows without experience in Hive, Pig, Java, Scala or Python to create new data integration jobs. You should be able to define a job without regard to the framework it will be executed in or whether it will be executed on-premises or in the cloud. The job design should be independent of the infrastructure

- Libraries of pre-built templates that allow you to offload mainframe data into modern platforms and easily replicate the same basic data transformations such as joins, sorts, aggregations etc. that were previously performed by batch COBOL and job control language (JCL) on the mainframe
- Lineage and metadata tracking for governance and compliance to various industry regulations, as well as to make the source data that drives analytics more transparent in its origins and transformations before analysis

- Automated performance optimization. Often programmers who write code spend as much or more time optimizing that code for performance. A good tool will automate this process, vastly shortening development time, and providing optimal performance on every hardware configuration or framework

Step Three: Minding the Security Gap

Mainframes manage some of the most sensitive data in the enterprise, and IT organizations are serious about protecting it. At the executive level, it's understood that companies cannot make mistakes when it comes to mainframe data security. So, what's the bottom line? If you cannot guarantee the secure access, processing and distribution of data, your Hadoop initiative won't make it out of the proposal phase.

As previously discussed, taking data off the mainframe is a hazy process, especially since you don't exactly know what you're getting until after you've retrieved it. This presents a huge security challenge. Taking data off the mainframe needs to be done in a secure manner, to ensure that only the data that is supposed to be accessed is accessed. This happens through a security infrastructure that integrates seamlessly into the mainframe environment.

Why Some Tools Don't Cut It

Some mainframe data extraction tools require a user to install unproven, untested software on the mainframe while others force you to adopt their own security model. In most enterprise IT organizations, neither of these approaches are used as a standard or allowed.

Best Practices to Close the Security Gap

Experience and trust are everything. Any solution you choose should be proven enterprise-grade software, with a strong security reputation.

Any solution that is used to accelerate the access, translation, transfer, and integration of mainframe data into Hadoop should include:

- No unnecessary installations on the mainframe
- Support for Kerberos and LDAP
- Secure data imports and exports
- Secure job execution
- User-level security while using an authorization protocol
- Direct and secure FTP support and Connect:Direct support
- Data encryption in flight

Step Four: Reducing the Cost Gap

As with all IT and big data projects, there's the cost factor. The costs of managing a terabyte of data can swing widely within the enterprise.

Given the extraordinary expenses of using MIPS on the mainframe, getting data off Big Iron and into the cloud or Hadoop cluster shouldn't incur any extra mainframe CPU cycles for sorting, filtering, copying or downloading. Otherwise, what's the point?

Best Practices to Close the Cost Gap

Any solution used to offload mainframe data and processing into a modern data platform should allow you to:

- Lower OPEX and CAPEX costs: Your company should be able to reduce operational expenses by minimizing CPU and I/O utilization on the mainframe, cutting down on expensive MIPS. You may also be able to save capital expenses by delaying upgrades to bigger and more expensive mainframes as your business grows. In addition, a good solution should provide automated performance optimization on the cluster. By optimizing the resource efficiency of the Hadoop cluster, you can maximize throughput per node, therefore reducing the need to constantly add more nodes.
- Gain faster time to insight: Your company should be able to process more data in less time while using the same resources. Your analysts should be able to make better, quicker decisions, based on more accurate insights from a complete picture of the company's data.
- Get the most from the cloud without silos, lock-in or rework as you move from on-premises to cloud or from one cloud to another.
- Cleanse, blend and transform data on the cluster: The solution should easily move mainframe data onto a Hadoop cluster without modifying it from its original format and make it distributable so you can store and process it on Hadoop. Or, you should be able to completely transform that data from the original mainframe format to a standard Hadoop format, and combine it with other enterprise data for richer context.
- You should be able to process mainframe variable data on the cluster without having to bloat the data out to its maximum record length, which would drastically bog down the processing speed.

Precisely Helps You Bridge All the Gaps

The **Precisely Connect** is designed to help you gain strategic value from your data, enabling you to access and integrate all your enterprise data connections whether on-premises or in the cloud.

Precisely Connect simplifies the process of connecting traditional systems – including mainframes – to downstream cloud applications, data lakes, and analytics platforms.

- Get mainframe data into cloud or big data platforms – keeping data in sync in real-time
- Take a design once, deploy anywhere approach using existing skills in your organization
- Guarantee data delivery for replication with no manual intervention
- Future-proof job designs for emerging frameworks

precisely

About Precisely

Precisely is the global leader in data integrity, providing accuracy and consistency in data for 12,000 customers in more than 100 countries, including 90 percent of the Fortune 100. Precisely's data integration, data quality, location intelligence, and data enrichment products power better business decisions to create better outcomes. Learn more at www.precisely.com.

www.precisely.com