databricks precisely

WHITEPAPER

A Practical Guide to Analytics and AI in the Cloud With Legacy Data

Databricks and Precisely: Access Legacy Data in the Cloud for Analytics, Data Science and Machine Learning

Contents

Introduction

Businesses that use legacy data sources such as mainframe and IBM i have invested heavily in building a reliable data platform. At the same time, these enterprises want to move data into the cloud for the latest in analytics, data science and machine learning. Databricks and Precisely have partnered to help make it easy and reliable to integrate legacy data into modern cloud data platforms for analytics and AI.

The Importance of Legacy Data

Mainframe and IBM i are still the processing backbone for many organizations, constantly generating important business data. It's crucial to consider the following:

MAINFRAME AND IBM i ARE THE ENTERPRISE TRANSACTION ENVIRONMENT

In 2019, there was a 55% increase in transaction volume on mainframe environments.[1] Additionally, IBM i serves as some of the most used systems for financial transactions, retail payments and logistical operations. Studies estimate that 2.5 billion transactions are run per day, per legacy system across the world.[2]

LEGACY IS THE FUEL BEHIND CUSTOMER EXPERIENCES

Within industries such as financial services and insurance, most customer information lives on legacy systems. Over 70% of enterprises say their customer-facing applications are completely or very reliant on mainframe processing.[3]

BUSINESS-CRITICAL APPLICATIONS RUN ON LEGACY SYSTEMS

Mainframe and IBM i systems often hold business-critical information and applications — from credit card transactions to claims processing. For over half of enterprises with a mainframe, they run more than half of their business-critical applications on the platform.[4]

However, they also present a limitation for an organization in its analytics and data science journey. While moving everything to the cloud may not be the answer, identifying ways in which you can start a legacy modernization process is crucial to the next generation of data and AI initiatives.

In this guide, we'll walk you through how Databricks and Precisely have come together to improve business insights by making legacy data accessible for modern analytics. Customers are turning to this joint solution to integrate mainframe and IBM i data to the Databricks Unified Data Analytics Platform in the cloud for their analytics, data science and machine learning use cases.

The Cost of Legacy Data

Across the enterprise, legacy systems such as mainframe and IBM i serve as a critical piece of infrastructure that is ripe with opportunity for integration with modern analytics platforms. If a modern analytics platform is only as good as the data fed into it, that means enterprises must include all data sources for success. However, many complexities can occur when organizations look to build the data integration pipelines between their modern analytics platform and legacy sources. As a result, the plans made to connect these two areas are often easier said than done.

DATA SILOS HINDER INNOVATION

Over 60% of IT professionals with legacy and modern technology in house are finding that data silos are negatively affecting their business. As data volumes increase, IT can no longer rely on current data integration approaches to solve their silo challenges[5].

CLOUDY BUSINESS INSIGHTS

Business demands that more decisions are driven by data. Still, few IT professionals who work with legacy systems feel they are successful in delivering data insights that reside outside their immediate department. Data-driven insights will be the key to competitive success. The inability to provide insights puts a business at risk.[6]

SKILLS GAP WIDENS

While it may be difficult to find skills for the latest technology, it's becoming even harder to find skills for legacy platforms. Enterprises have only replaced 37% of the mainframe workforce lost over the past five years. As a result, the knowledge needed to integrate mainframe data into analytics platforms is disappearing.[7]

While the drive for building a modern analytics platform is more powerful than ever, taking this initiative and improving data integration practices that encompass all enterprise data has never been more challenging. The success of building a modern analytics platform hinges on understanding the common challenges of integrating legacy data sources and choosing the right technologies that can scale with the changing needs of your organization.

Challenges Specific to Extracting Mainframe and IBM i Data

With so much valuable data on mainframe and IBM i, the most logical thing to do would be to connect these legacy data sources to a modern data platform. However, many complexities can occur when organizations begin to build integration pipelines to legacy sources. As a result, the plans made to connect these two areas are often easier said than done. Shared challenges of extracting mainframe and IBM i data for integration with modern analytics platforms include the following:

DATA STRUCTURE

It's common for legacy data not to be readily compatible with downstream analytics platforms, open-source frameworks and data formats. The varied structures of legacy data sources differ from relational data. Legacy data sources have traits such as hierarchical tables, embedded headers, and trailer and complex data structures (e.g., nested, repeated or redefined elements). With the incorrect COBOL redefines and logic set up at the start of a data integration workflow, legacy data structures risk slowing down processing speeds to the point of business disruption and can lead to incorrect data for downstream consumption.

METADATA

COBOL copybooks can be a massive hurdle to overcome for integrating mainframe data. COBOL copybooks are the metadata blocks that define the physical layout of data but are stored separately from that data. As a result, they can be quite complicated, containing not just formatting information, but also logic in the form, for example, of nested Occurs Depending On clauses. For many mainframe files, hundreds of copybooks may map to a single file. Feeding mainframe data directly into an analytics platform can result in significant data confusion.

DATA MAPPING

Unlike an RDBMS, which needs data to be entered into a table or column, nothing enforces a set data structure on the mainframe. COBOL copybooks are incredibly flexible so that they can group multiple pieces into one, or subdivide a field into various fields, or ignore whole sections of a record. As a result, data mapping issues will arise. The copybooks reflect the needs of the program, not the needs of a data-driven view.

DIFFERENT STORAGE FORMATS

Often numeric values stored one way on a mainframe are stored differently when the data is moving to the cloud. Additionally, mainframes and IBM i use a whole different encoding scheme (EBCDIC vs. ASCII) — it's an 8-bit structure vs. a 7-bit structure. As a result, multiple numeric encoding schemes allow for the ability to "pack" numbers into less storage (e.g., packed decimal) space. In addition to complex storage formats, there are techniques to use each individual bit to store data.

Whether it's a lack of internal knowledge on how to handle legacy data or a rigid data framework, ignoring legacy data when building a modern data analytics platform means missing valuable information that can enhance any analytics project.

Pain Points of Building a Modern Analytics Platform

Tackling the challenges of mainframe and IBM i data integration is no simple task. Besides determining the best approach for integrating these legacy data sources, IT departments are also dealing with the everyday challenges of running a department. Regardless of the size of an organization, there are daily struggles everyone faces, from siloed data to lack of IT skills.

ENVIRONMENT COMPLEXITY

Many organizations have adopted hybrid and multi-cloud strategies to manage data proliferation, gain flexibility, reduce costs and increase capacities. Cloud storage and the lakehouse architecture offer new ways to manage and store data. However, organizations still need to maintain and integrate their mainframes and other on-premises systems — resulting in a challenging integration strategy that must encompass a variety of environments.

SILOED DATA

The increase in data silos adds further complexity to growing data volumes. Data silo creation happens as a direct result of increasing data sources. Research has shown that data silos have directly inhibited the success of analytics and machine learning projects.

PERFORMANCE

Processing the requirements of growing data volumes can cause a slowdown in a data stream. Loading hundreds, or even thousands, of database tables into a big data platform — combined with an inefficient use of system resources — can create a data bottleneck that hampers the performance of data integration pipelines.

DATA QUALITY

Industry studies have shown that up to 90% of a data scientist's time is getting data to the right condition for use in analytics. In other words, 90% of the time, data feeding analytics cannot be trusted. Data quality processes that include mapping, matching, linking, merging, deduplication and actionable data are critical to providing frameworks with trusted data.

DATA TYPES AND FORMATS

Valuable data for analytics comes from a range of sources across the organization from CRM, ERPs, mainframes and online transaction processing systems. However, as organizations rely on more systems, the data types and formats continue to grow. IT now has the challenge of making big data, NoSQL and unstructured data all readable for downstream analytics solutions.

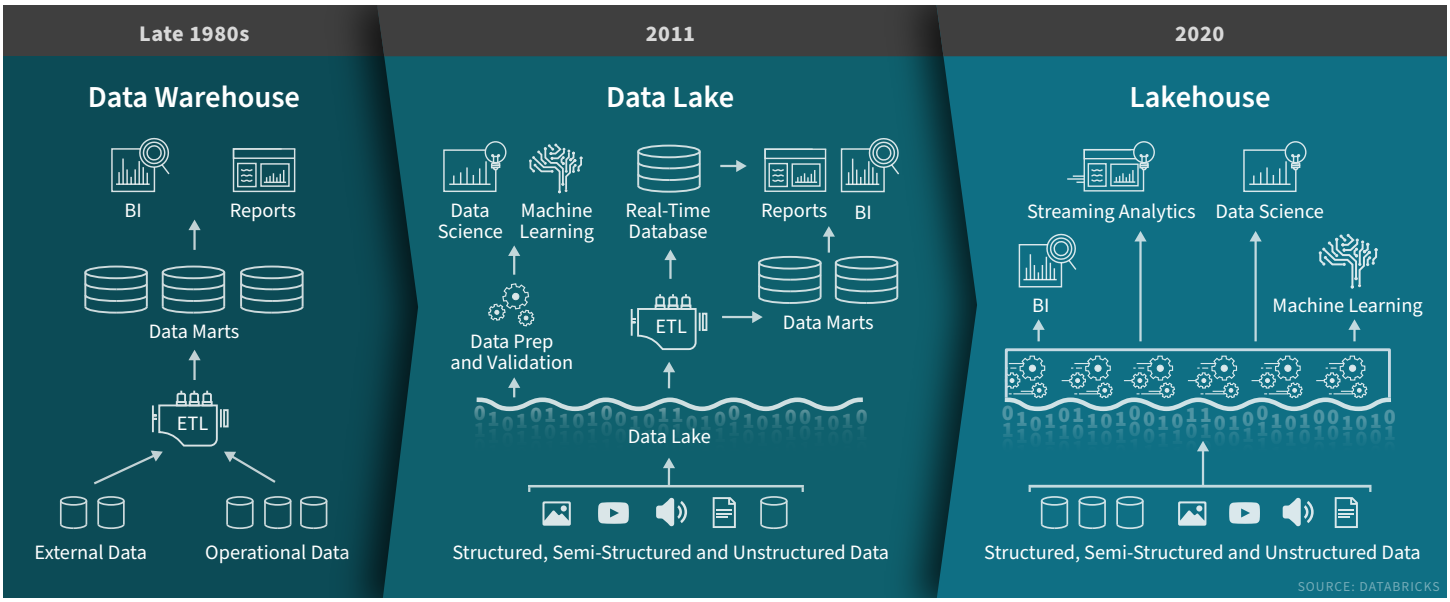
SKILLS GAP AND RESOURCES

The need for workers who understand how to build data integration frameworks for mainframe, IBM i, cloud, and cluster data sources is increasing, but the market cannot keep up. Studies have shown that unfilled data engineer jobs and data scientist jobs have increased 12x in the past year alone. As a result, IT needs to figure out how to integrate data for analytics with the skills they have internally.

What Your Cloud Data Platform Needs

A new data management paradigm has emerged that combines the best elements of data lakes and data warehouses, enabling analytics, data science and machine learning on all your business data: lakehouse. As described in the Databricks blog, What is a lakehouse:

> *Lakehouses are enabled by a new system design: implementing similar data structures and data management features to those in a data warehouse, directly on the kind of low-cost storage used for data lakes. They are what you would get if you had to redesign data warehouses in the modern world, now that cheap and highly reliable storage (in the form of object stores) are available.*



This new paradigm is the vision for data management that provides the best architecture for modern analytics and AI. It will help organizations capture data from hundreds of sources, including legacy systems, and make that data available and ready for analytics, data science and machine learning. A lakehouse has the following key features:

- **Open storage formats**, such as Parquet, avoid lock-in and provide accessibility to the widest variety of analytics tools and applications
- **Decoupled storage and compute** provides the ability to scale to many concurrent users by adding compute clusters that all access the same storage cluster
- **Transaction support** handles failure scenarios and provides consistency when multiple jobs concurrently read and write data
- **Schema management** enforces the expected schema when needed and handles evolving schemas as they change over time
- **Business intelligence** tools directly access the lakehouse to query data, enabling access to the latest data without the cost and complexity of replicating data across a data lake and a data warehouse
- **Data science and machine learning** tools used for advanced analytics rely on the same data repository
- **First-class support for all data types** across structured, semi-structured and unstructured, plus batch and streaming data

Delta Lake

Delta Lake is an open-source storage layer that brings the reliability, performance and lifecycle management to data lakes as critical components of any modern data architecture such as a lakehouse. This has emerged as the new, open standard for data lakes that runs on top of your existing data lake and is fully compatible with Apache Spark™ APIs.

ACID TRANSACTIONS

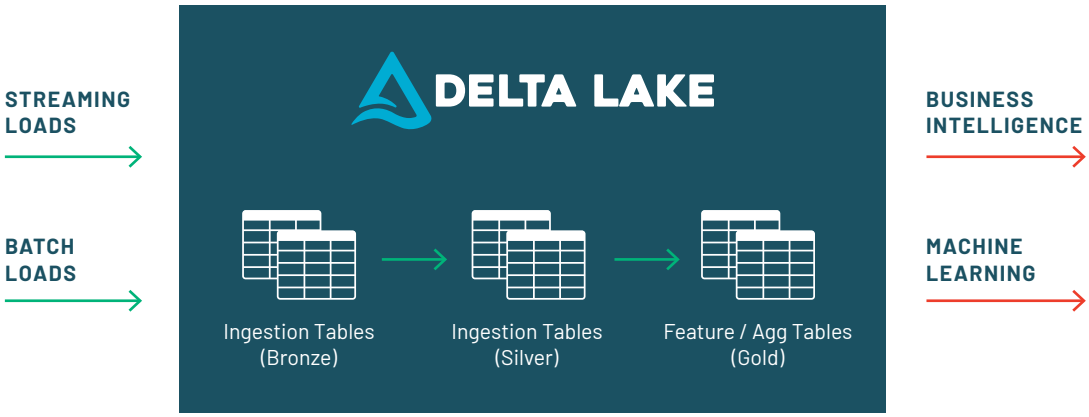
Delta Lake is implemented using open-format Parquet files with a transaction log that provides ACID transaction support and defines what data and files are the most recent so when a job queries the data sets, users are presented with accurate, consistent data sets. Other key features include scalable metadata handling, time travel (data versioning), schema enforcement, schema evolution and audit history.

PROCESSING UPDATES

One historical problem with data lakes has been the immutable nature of common file formats used to store data in data lakes, which made it difficult to update existing data sets. This often meant extensive and complex development was required to build data pipelines that could process updates and deal with failure scenarios. A favorite capability of Delta Lake addresses this challenge as well, by providing native functionality to update and merge data with existing data sets.

BRONZE, SILVER, GOLD DATA

The following diagram outlines the best practice of using Delta Lake to ingest and store raw data (Bronze) from any source into your data lake, then process that data into filtered and cleansed tables (Silver), and finally into aggregated data (Gold) ready for downstream consumption. By processing data in these stages, the data quality and readiness is improved at each step, making it ready for analytics while keeping the raw, detailed data available for analysis whenever needed.



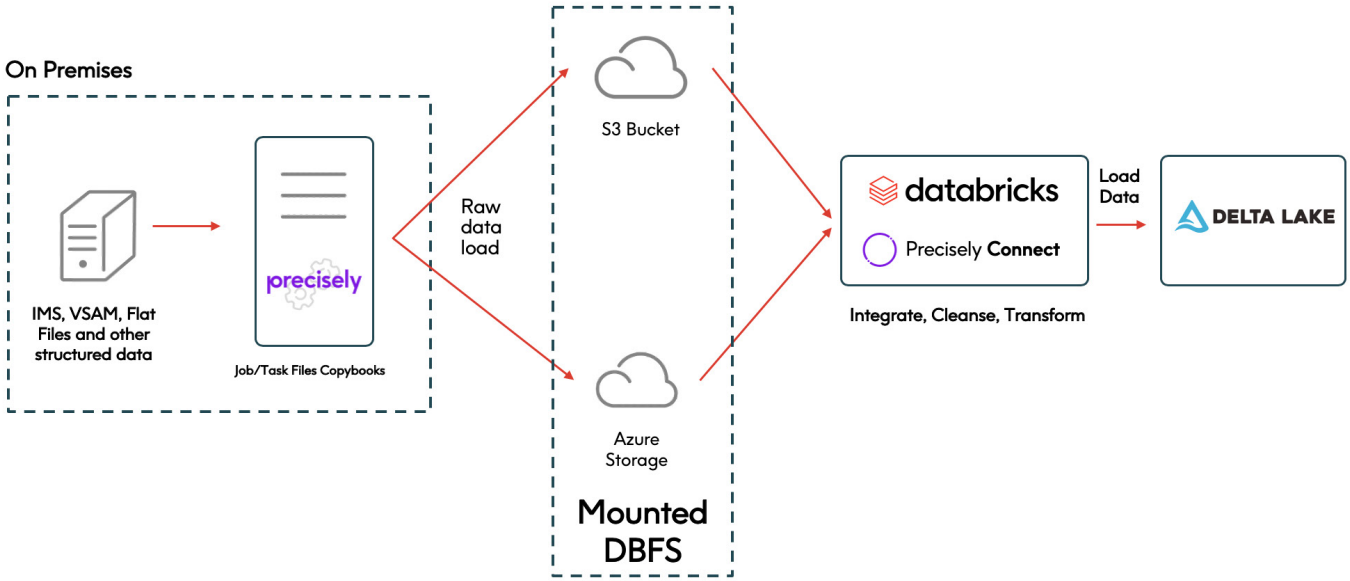
How Databricks and Precisely Integrate Mainframe Data Into Delta Lake on Databricks

STEP 1: Establish the types of mainframe data assets that need to be made accessible within Databricks. These can be fixed length and sequential flat files, VSAM files, Db2 tables or IMS databases. For flat files and VSAM files, identify the copybooks that store the mainframe metadata necessary for proper data conversion and translation.

STEP 2: Design a visual ETL pipeline to convert the data based on the structure laid out in the copybook. Interpret records in a variety of ways based on rules in the copybook. Often the data needs to be broken apart from its hierarchical format on the mainframe and mapped to a relational model to easily load into Delta Lake.

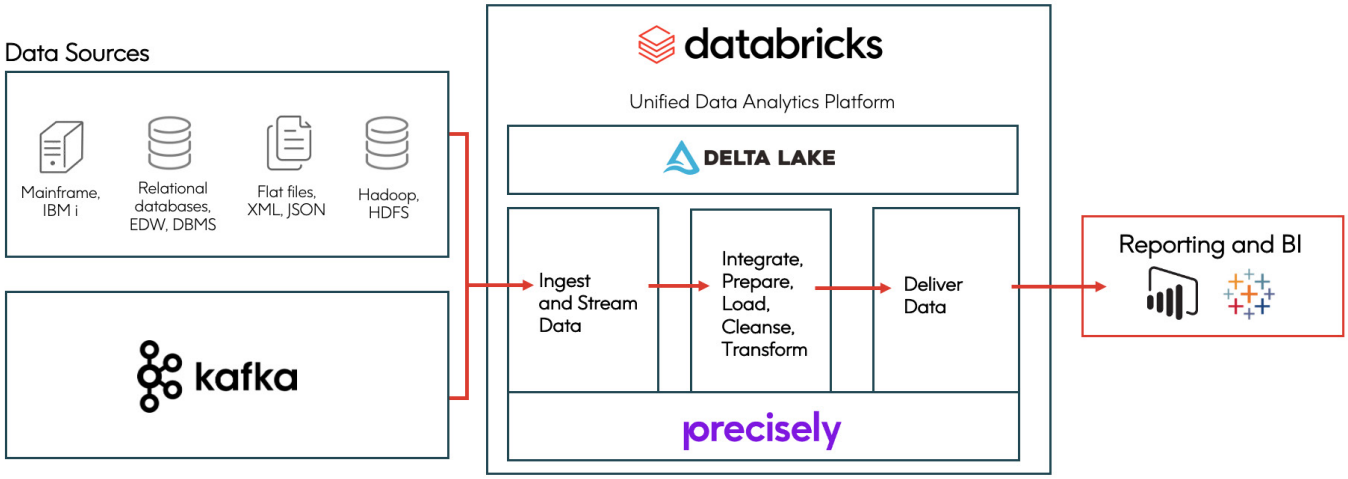
STEP 3: Deploy the pipeline. The Precisely Connect design-once, deploy-anywhere architecture ensures that data can be loaded into Delta Lake from a single compute instance. Additionally, using Precisely, users can leverage the distributed processing power of Databricks to scale data ingestion and conversion across the compute cluster.

STEP 4: Ensure the latest changes are being delivered to Databricks by leveraging the CDC capabilities of Connect for Db2/z, VSAM and IMS. Only changed data is delivered, ensuring minimal system impact and maximum scalability as transactional volumes continue to grow. Analytics and AI applications running on Databricks will always be operating off the latest data.



Making It Tangible: Insurance Company Builds Claims Data Hub With Databricks and Precisely

Building a modern analytics platform means understanding how to unlock the value of legacy data. Now, let's take a look at how one American insurance company has successfully created an enterprise-wide claims data hub. This modern analytics platform includes both new and legacy data sources.



THE GOAL

An American insurance company wanted to take a variety of data from across their organization to build an enterprise-wide claims data lake. The purpose of the claims data lake was to receive data from across the lines of business and improve analysis of customer activity, historical data and richer analytics. In its ideal scenario, the claims data would help identify patterns in claims to alert the business to unexpected severe claims or to automate the fast-tracking of low dollar claims without the need for an adjuster.

THE CHALLENGE

Data funneling into the hub would include information from core systems such as actuary, call center, claims and billing for different departments. Most of this data existed on mainframes. Mainframe data file formats included EBCDIC-encoded VSAM data with binary and packed data types mapped by multiple complex copybooks. When it came time to integrate all these data sources, the insurance company struggled to get data from the mainframe to its data lake. Getting mainframe data into the data lake meant that they had to spin up an entirely separate process for data ingestion. As a result, the insurance company had a siloed process that caused lost time, delayed delivery and incomplete claims analytics.

THE SOLUTION

1. Ingestion and streaming of legacy data

The insurance company knew that the first problem they needed to tackle was to break down the silos that existed within their legacy data sources. Using Precisely Connect, the insurance company was able to ingest mainframe data sources and make them readable for use within the Databricks Unified Data Analytics Platform. Connect was able to perform this task due to its ability to access and understand VSAM, mainframe fixed and variable files, and Db2 data directly. Also, Precisely helped the insurance company to leverage its existing mainframe metadata in the ingestion process by mapping to the existing COBOL copybooks on the mainframe. The insurance company, like many organizations, did not have mainframe skills in-house but was able to perform the integration and streaming of the data due to Precisely 's low-code approach.

2. Integration and preparation of legacy data — HDInsights offload

Once mainframe data ingest was complete, the insurance company then needed to modernize its ETL processes to scale within Databricks. The insurance company had been using Connect with Spark on Azure HDInsights for ETL transformation on its claims data hub data and determined a need to move these existing workflows into Databricks. However, the insurance company did not want to perform any rework to their data integration workflows, especially as many had complex data transformations upon the mainframe data.

Using Connect, the insurance company built ETL processes that took a design-once, deploy-anywhere approach and as a result, had no rework or redesigns required to migrate the Azure HDInsights pipelines to run on Databricks. Data migration from Hive on HDInsights to Delta Lake was achieved via JDBC connectivity and the Connect high-performance integration engine to sufficiently parallelize the data load. Furthermore, Connect was able to produce the high-performance, self-tuning sorts, joins, aggregation, merges and lookups required for the organization to get the data they needed in the right way. Connect's ability to run natively in the Databricks runtime also ensured they were able to optimize the data integration workflow for the high-volume requirements of the claims data hub.

3. Delivering data changes with Databricks and Precisely

To ensure that the data fed into the claims data hub was always up to date, the insurance company also needed a way to capture data changes off the mainframe in real time. Connect helped the insurance company deliver mainframe data changes in real time through efficient log-based capture from legacy stores — VSAM, Db2/z — to Databricks. Additionally, the insurance company chose to leverage Connect for its resilient data replication to ensure data integrity in case of a connection failure between the mainframe and Databricks.

Conclusion

Together, Precisely and Databricks eliminate data silos across the business to get high-value, high-impact, and complex data ready for analytics, data science and machine learning in the cloud. By building out a data integration framework and cloud data platform that bring legacy data into Databricks, organizations can maximize the value of their advanced analytics.

Furthermore, organizations need to assess the data integration frameworks that they depend on to bring them the comprehensive analytics required for competitive survival. Without setting in place technologies that enable seamless connection of sources/targets and the scalability to process increasing data volumes, there's a risk of being left behind.

---

[1] The Next Wave of Mainframe Success, BMC 2019

[2] SHARE IBM User Group, 2017

[3,4] Forrester Consulting, 2018

[5] *Data Trends for 2019: Extracting Value from Data*, Precisely 2019, N = 58 respondents who own a mainframe

[6] *Data Trends for 2019: Extracting Value from Data*, Precisely 2019, N = 58 respondents who own a mainframe

[7] Forrester Consulting, 2018