WHITEPAPER

# Enterprise Data Pipelining

## The Guide to Effective Data Preparation and Delivery

This paper responds to “What is a data pipeline?” by discussing its definition, use cases, challenges, benefits, and a new approach based on data products.

# Make your big data stores analytics-ready

**Today's enterprise needs access to high- quality big data for analytical and operational use cases.**

An enterprise data pipeline is critical to the success of all types of data processing, whether offline or real- time, analytical or operational.

Here are just a few enterprise use cases where a consistent, accurate data pipeline is critical:

- Customer 360
- Data Masking
- Test Data Management
- Data Migration
- Legacy Application Modernization

# What is a Data Pipeline?

**A data pipeline is a means of moving data from its source, to its target destination (such as a data lake or data warehouse). On the way, the data is transformed and enriched, so that it can be analyzed, and used to generate business insights**

More specifically, a data pipeline represents the steps involved in collecting, joining, culling, cleansing, and otherwise transforming data into a form that applications and users can trust, and readily ingest, for various use cases.

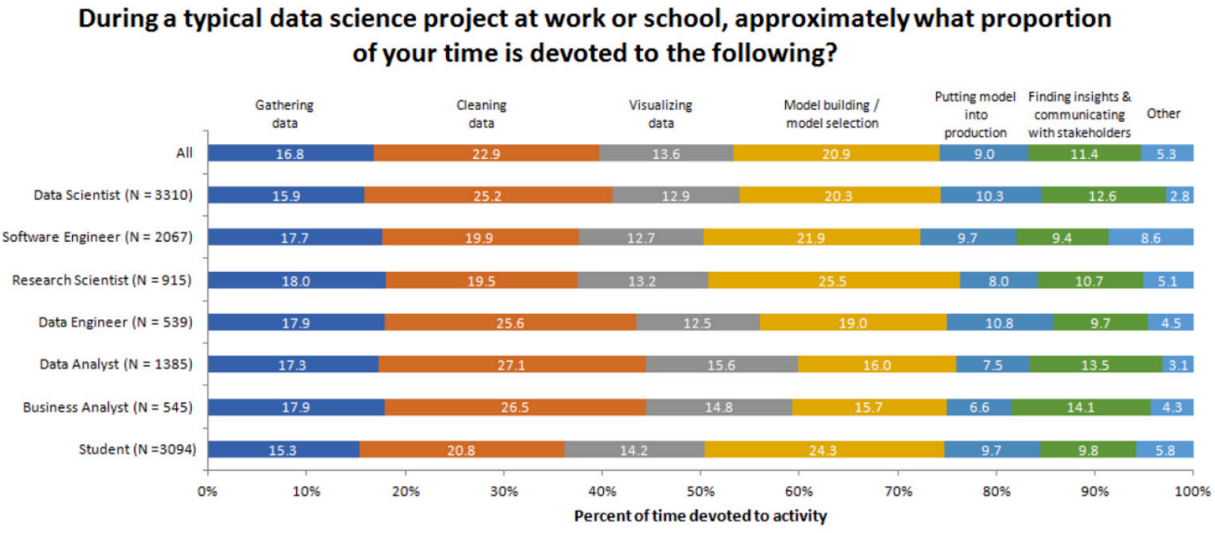
To begin to appreciate the importance of a data pipeline, you need to understand that data analysts and data scientists are involved with:

1. **Building** a data pipeline with data pipeline tools
2. **Visualizing** the data, via a data catalog
3. **Selecting** or creating the right models
4. **Deploying** the models into production
5. **Extracting** insights, with an operational intelligence platform
6. **Operationalizing** data flows

Now we can better understand the market need.

# Why is a Data Pipeline Necessary?

A Kaggle survey of 23,000 data professionals revealed that data preparation accounts for 44% of a data analyst's time, and 41% of a data scientist's time. Only 11% of their time is spent extracting insights from data – which is their primary job.




*Source: Business Over Broadway*

Time spent on data preparation is constantly growing across the enterprise in support of offline big data analysis (by data analysts and data scientists), as well as online use cases, supported by an operational intelligence platform. The need for trusted, high-quality, complete, and current data – to drive these use cases – is paramount. As a result, slow, tedious, error-prone, and manual processes for data preparation are simply no longer acceptable.

What is a data pipeline, if not to ensure confidence. Confidence in the accuracy of the data, the way it is prepared, and of course, the results and insights derived from it. A consistent data pipeline ensures that consuming applications and analysis will perform as expected, and that the resultant insights can be trusted time after time. A poorly constructed data pipeline can lead to the wrong conclusions – that is to say, "garbage in, garbage out."

## Data Pipeline Use Cases

**Accessing complete, accurate data for both analytical and operational workloads is difficult to do. Today's enterprises need a single, self-service, automated data pipeline, to serve both offline and online use cases.**

### Offline use cases

Offline use cases typically rely on well-prepared data that is stored in data warehouses or data lakes. As they analyze historical data, data analysts and scientists aren't concerned with up-to-date data. But the sheer size of these mega data stores means that data preparation and the data pipeline can take a long time. Common examples requiring high-quality, offline data include:

- Ad-hoc business intelligence
- Predictive analytics
- Machine learning
- Test data provisioning
- Customer Data Platform, or Hub, for marketing use cases

### Online use cases

Unlike offline analytics, online, or operational, use cases need up-to-date data to conduct rapid analysis, and to provide actionable insights – automatically and in real time. Online use cases require access to fresh data, generated via real-time data integration tools in sync with live operational systems, not stale historical data from a data warehouse. Here are some examples:

- Customer 360, for call centers and customer self-service portals
- Next-best-action recommendations (for example, customer churn)
- Product recommendations
- Real-time fraud detection
- Credit scoring and approvals/rejections
- Network monitoring and intrusion detection
- Artificial intelligence and automated real-time decisioning

# 7 Steps to a Working Data Pipeline

We know why consistent, accurate data is essential, but why does preparing it take so long? We need only to look at the multitude of steps involved, to see why:

1. **Data collection -** Identifying the data sources, target locations for backup/storage, frequency of collection, and setting up/ initiating the mechanisms for data collection.
2. **Data discovery and profiling -** Pinpointing the data of interest, reviewing and understanding source data structure, interrelationships, and locating the systems or databases in which desired data resides.
3. **Data cleansing -** Detecting corrupt, missing, outlier, inaccurate, or irrelevant data elements in a dataset, then replacing, modifying, or removing them, so they won't adversely affect the outcome.
4. **Data structuring -** Defining the target structure in which to store the prepared data. The original forms are rarely the best fit for the desired analysis, or any other use case.
5. **Data transformation, enrichment, and masking –** Matching the format of the source data to the target structure, by providing missing or incomplete information, and employing data masking tools, to protect sensitive information from unauthorized access and use.
6. **Data validation -** Checking the quality and accuracy of the data to make sure it's suitable for the target application/consumer.
7. **Data delivery -** Delivering the prepared data to the target applications, real-time decisioning engines, data lakes, or data warehouses.

Only after completing these 7 steps, can you have confidence in the insights you receive.

# Data Pipeline Challenges

The aforementioned steps are required to get data from different internal systems and external sources into a form usable by the target application. Even if the steps are scripted and semi- automated for offline use cases, they have to be scheduled and initiated well in advance.

On the other hand, operational use cases generally require complex real-time integrations to dozens of enterprise systems, as well as coding of all these steps, for real-time execution.

But that's just the beginning. There are 7 data management challenges that make an enterprise data pipeline much more difficult to achieve:

- **Technology and structure**
  Each data source has its application- or domain-centric view of the world, and many of them are built in different technologies, and different structures.
- **Consistent data integration**
  Accessing and joining related data from these systems can be an integration nightmare, to say nothing about keeping the data consistent across them.
- **Refresh/restart interruptions**
  Every time you refresh your data, you have to start the data pipeline steps from the beginning, meaning that your analytics, or product testing, grinds to a halt.
- **Error correction**
  Data is commonly created with missing values, inaccuracies, or other errors, that need to be corrected.
- **Format reconciliation**
  Different data sets often have different formats, that must be reconciled.
- **Quality verification**
  Verifying data quality, assuring compliance, and joining datasets constitute a big part of the data pipeline construction process.
- **Data transformation**
  Both offline and operational use cases require raw data to be accurately transformed, in order to be used effectively.

# Data Pipeline Benefits

If the enterprise could automatically access all the quality, up-to-date data it needs for any use case that requires big data preparation, how much different would the business landscape be? Having consistent, accurate, and timely data available for both analytical and operational use cases would make things better in 6 key ways:

**Better use of resources**
Data scientists can spend more time analyzing data, instead of finding and prepping it.

**Better results and insights**
Pipelining higher quality data results in more reliable results from analytics and other applications – quality in, quality out.

**Better consistency across use cases**
The ability to reuse a dataset, for more than one application, saves time and resources.

**Better consistency and integrity**
Automatic cleansing, transformation, masking, and every other step in the data preparation process, means no human error.

**Better Insights**
More accurate analysis leads to more reliable business insights, and outcomes.

**Better ROI**
A more consistent and automated data pipeline translates into faster ROI from the tools and applications that use the data.

# Data Pipelines via Data Products

The enterprise needs access to data for both offline (analytical) and online (operational) use cases, but lacks easy-to-use, self-service data pipelining tools that can quickly and consistently prepare and deliver high-quality data. That's because most organization are used to preparing their data, database by database, table by table, and row by row, joined with other data tables through a vast array of joins, indexes, and scripts.

For example, all invoices, all payments, all customer orders, all service tickets, all marketing campaigns, and so on, must be collected and processed separately. Then complex data matching, validation and cleansing logic is required to ensure that the data is complete and clean – with referential integrity maintained at all times.

A better way is to collect, cleanse, transform, and mask data is holistically, via data products. For example, a customer data product includes all master data, as well as interaction data and transaction data (e.g.,invoices, payments, service tickets, and campaigns.

# Building an Enterprise Data Pipeline on a Data Product Platform

Data Product Platform allows you to define a schema to accommodate all the attributes you want to capture for a given business entity (such as a customer or order), regardless of the source systems in which these attributes are stored. It automatically discovers where the desired data lives across all enterprise systems. It then collects the data from the source systems and stores it as an individually encrypted Micro-Database™, ready to be delivered to any authorized data consumer, on demand.

The platform keeps the data synchronized with the sources on a schedule you define. It automatically applies filters, transformations, enrichments, masking, and other steps crucial to a high-quality data pipeline. So, your data is always complete, up-to-date, and consistently and accurately prepared, ready for analytics or any operational use case you can conceive of.

Looking at your data holistically, at the data product level, ensures data integrity by design, giving your data teams quick, easy, and consistent access to the datasets they need. You always get insights you can trust, because you have data you can trust.

## Conclusion:

## Operationalize Data, Accelerate Time to Insights

Data scientists are rare and expensive resources in all organizations. To maximize their skills and experience, they should focus on building data models, and running predictive analytics – and not on understanding the structure of the data, and where to find it.

They need their data pipeline tools to be fully automated, without compromising their abilities to explore the data, and run experimental models.

Today, most self-service data preparation tools deliver partial automation for ad-hoc data preparation tasks. They can't operationalize data, in order to achieve a fully automated data pipeline, whose flows can be built once, and reused afterwards.

Data flows are iterative, and normally built, tested, and packaged by data engineers. Once the flows are packaged, they can be invoked and scheduled by data scientists.

When activating a pre-built data preparation flow, the data scientist can decide which data to generate (e.g., where, when, and how it will be delivered).

An Enterprise Data Pipeline uses data versioning, so data scientists can also reproduce previous data sets, as well as access historical versions of the data.
It keeps your data lakes and data warehouses in sync with your data sources, based on pre-defined data sync rules. Data changes can be ingested into your data stores via the delivery method of your choice: Extract, Transform, Load (ETL), data streaming, Change Data Capture (CDC), or messaging.

## About K2View

At K2View, we believe that every enterprise should be able to use its data to be as disruptive and agile as the best companies in its industry. We make this possible by enabling data teams to transform fragmented data into complete and compliant data products – in real time.

Data products could be customers, products, suppliers, orders – or anything else that's important to your business. We manage every individual data product in its own secure Micro-Database™, continuously syncing it with all source systems, and making it instantly accessible to authorized data consumers

This is all made possible by our Data Product Platform, which delivers a trusted, real-time view of anybusiness entity. The platform deploys in weeks, scales linearly, and adapts to change on the fly.
It supports modern data architectures, such as data mesh, data hub, and multi-domain MDM
– in on-premise, cloud, or hybrid environments.

This one platform drives many use cases, including application modernization, cloud migration, customer 360, data privacy, data testing, and more – to deliver business outcomes in less than half the time, and at half the cost, of any other alternative.