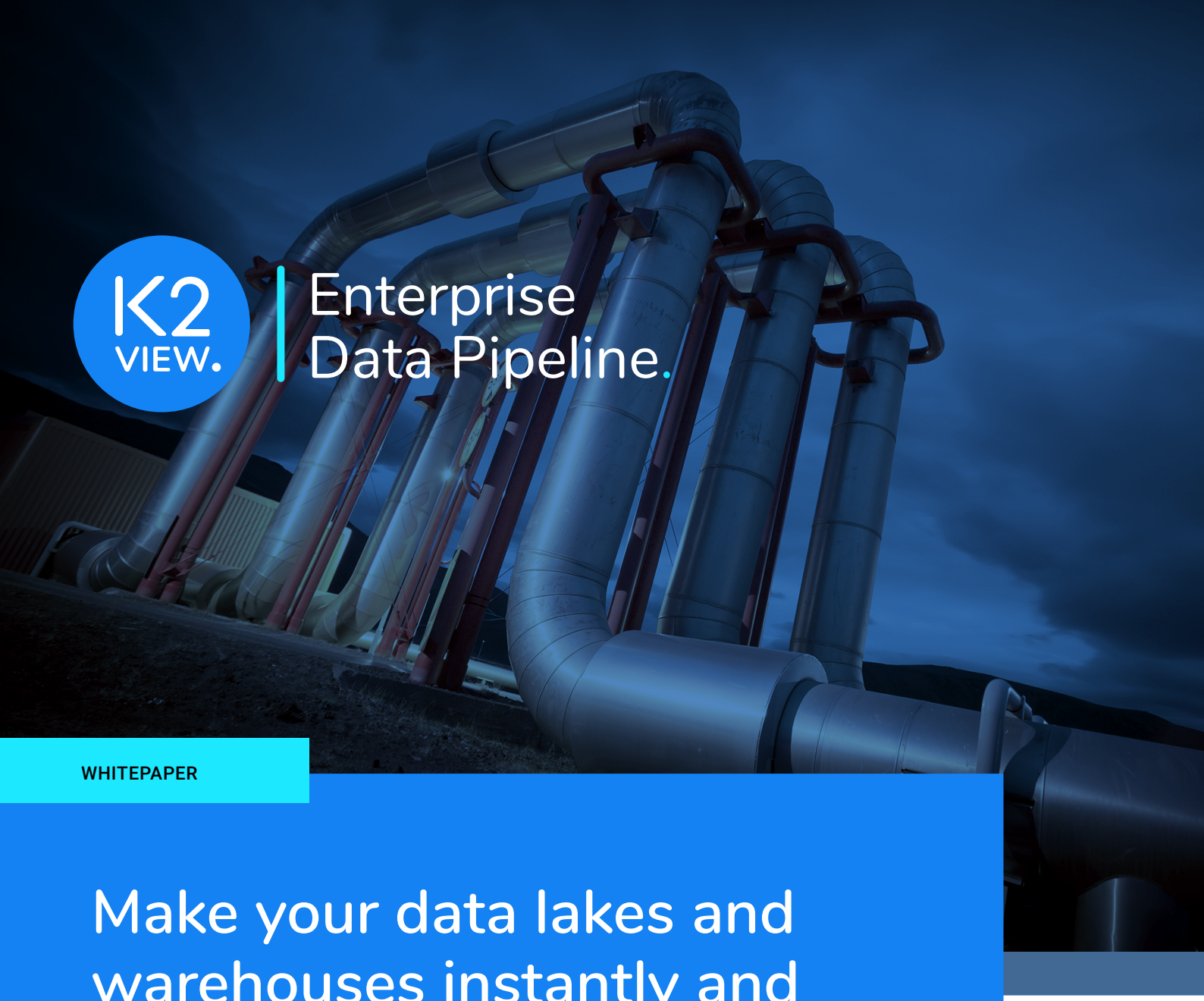


K2 VIEW.

Enterprise Data Pipeline.

WHITEPAPER

Make your data lakes and warehouses instantly and always ready for analytics

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

Data preparation and pipeline is critical to the success of all types of data processing, whether offline or real-time, and whether they are analytical or operational in nature. Here are just a few enterprise use cases where consistent, accurate data preparation and pipeline is critical.

Introduction

Data professionals spend their time in many different activities during their data analysis assignments -

- Preparing and pipelining the data
- Visualizing the data
- Building or selecting the right models
- Depolying the models into production
- Extracting insights
- Operationalize data flows

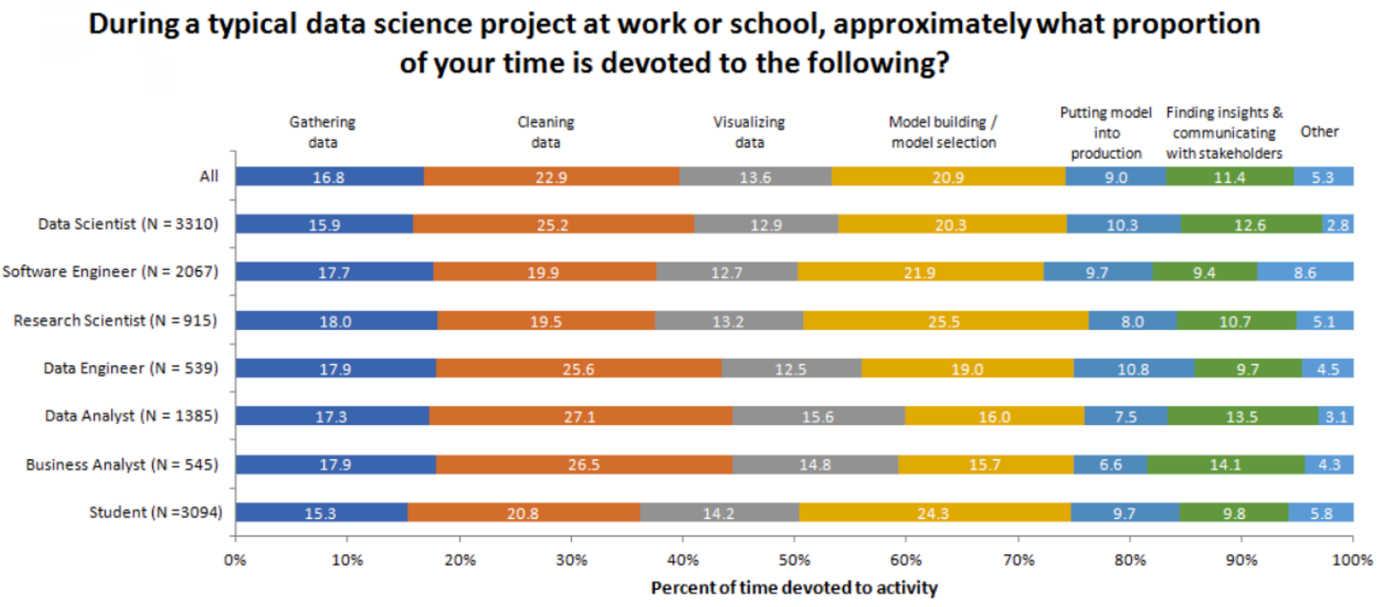
A recent survey of 23,000 data professionals revealed that data preparation accounts for 44% of a data analyst's work time, and 41% of a data scientist's time. Only 11% of their time is spent extracting insights from data – which is their primary role.

> Data pipeline is the process of collecting, joining, culling, cleansing, and otherwise transforming big data into a form that applications and users can trust and readily ingest for various use cases.

Time spent on data preparation is constantly growing across the enterprise in support ofofflinebig data analysis(by data analysts and data scientists), as well as online operational use cases, in support ofreal-time intelligence. 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 and pipeline are simply no longer acceptable.

Why is proper data preparation and pipeline so necessary? Simply put, it's to ensure confidence. Confidence in the accuracy of the data, the way it isprepared, and of course, the results and insights we get from it. Consistent data preparation ensures that consuming applications and analysis will perform as expected and that the insights derived can be trusted every time. Poorly prepared data can lead to erroneous conclusions - as they say, "garbage in, garbage out."

Today's organizations seek a single platform for self-service, automated data preparation and pipeline, for all enterprise use cases - both online and offline.



Offline and operational use cases

Data preparation and pipeline is critical to the success of all types of data processing, whether offline or real-time, and whether they are analytical or operational in nature. Here are just a few enterprise use cases where consistent, accurate data preparation and pipeline is critical.

Offline use cases

These typically rely on well-prepared data that is stored in data warehouses and data lakes. As they analyze and operate on historical data, they aren't concerned with up-to-the-moment data. However, the massive size of these data stores means data preparation and pipeline can take along time.

Some common examples that require high-quality offline data include:

- Ad-hoc business intelligence
- Predictive analytics
- Machine learning
- Test data provisioning
- Customer Data Platformfor marketing use cases

Operational intelligence

Unlike offline analytics, operational use cases need up-to-date data to make rapid analyses, provideactionable in-the-moment insights,provide better customer experiences, and automate decisions - in real-time and based on the latest data. That means operational intelligence use casesrequire data integration to live operational systems, not just static historical data from a big data warehouse. Here are some examples:

- Customer 360 for call center and customer self-service
- 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 of data pipelining

We know why consistent, accurate data is essential, but why does preparing it take so long? We need only 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 - Identifying the data of interest, reviewing and understanding source data structure, interrelationships, andlocating system(s) or database(s) wherethe desired data reside.
3. Data cleansing - Detectingcorrupt, 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 – Changing the structure/format of the source data to match the target structure, provide missing or incomplete values, and changing values to protect sensitive information from unauthorized access and use.
6. Data validation - Checking the quality and accuracy of the data produced by all the previous steps to ensure it is 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 all these steps are completed, can you consume and trust the resulting data - confident in the insights you will receive.

Doing it right can be difficult

The above steps are required to get data from different internal systems and external sources into a formusable 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 rapid in-the-moment execution.

But that's just the beginning. There are other data management-related problems that make data preparation and pipeline that much harder.

- 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.
- Accessing and joining related data from these systems can be an integration nightmare, to say nothing about keeping the data consistent across them.
- Each time you need to refresh the data, you must restart the data preparation and pipeline steps, meaning your analytics or product testing grinds to a halt.
- Data is commonly created with missing values, inaccuracies, or other errors.
- Separate data sets often have different formats that you must reconcile.
- Correcting data errors, verifying data quality, masking, and joining datasets constitutes a big part of the data pipeline preparation process.
- Both offline and operational use cases require raw data tobe consistently and accurately prepped so that the results will also be consistent and valid.

The benefits of trusted data.

If the enterprise could **automatically** access all the quality, up-to-date data it needs for any use case that requires big data preparation pipeline, 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 so many ways:

1. **Better use of resources –** Data scientists can spend far more time analyzing data instead of finding and prepping it.
2. **Better results and insights –** Higher quality data as input means more reliable results from analytics and other applications—quality in, quality out.
3. **Better consistency across use cases –** The ability to reuse a single source of prepared data for more than one application saves time and resources.
4. **Better consistency and integrity –** Automatic cleansing, transformation, masking,and every other step in data preparation and pipeline means no human error.
5. **Better insights** mean better decisions.
6. **Better ROI -** Consistent and automatic data preparation and pipeline means faster ROI from the tools and applications that use the data.

Making data pipeline easy, foolproof, and fast – business entities

The enterprise needs a single data pipeline solution for offline and real-time operational use cases. The biggest obstacle to realizing these benefits is the lack of simple-to-use, self-service tools that can quickly and consistently pipeline quality data. That's because most big data is prepared database by database, table by table, and row by row, joined with other data tables through 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 maskdata as a business entity – holistically. So, for example, a customer business entity would include a single customer'smaster data, together with their invoices, payments, service tickets, and campaigns.

K2View Data Fabric supports data preparation by business entities. It allows you to define a Digital Entity schema to accommodate all the attributes you want to capture for a given business entity (like a customer or an order), regardless of the source systems in which these attributes are stored. It automatically discovers where the desired data across your enterprise systems, and it creates data connections to those source systems. K2View Data Fabric collects data from source systems and stores it as an individual digital business entity, each in a separate encrypted micro-database, ready to be delivered to a consuming application, big data store, or user.

The solution keeps the data synchronized with the sources on a schedule you define. It automatically applies filters, transformations, enrichments, masking, and other steps crucial to quality data preparation. 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.

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

Why you should pipeline data by business entity

Operationalize data to 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.

Data scientists need the data preparation process to be full automated, without compromising their abilities to explore the data, and run experimental models.

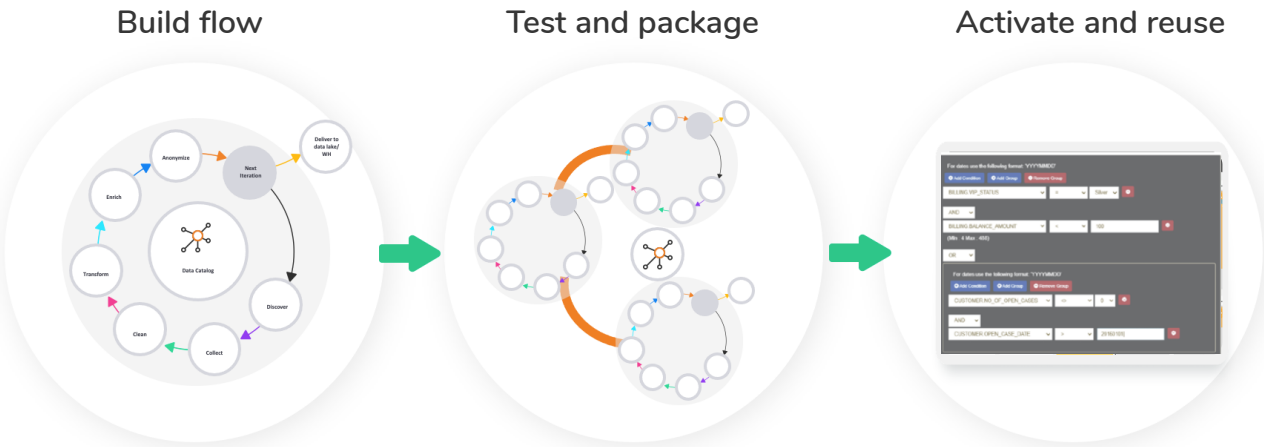
Today, most self-service data preparation tools deliver partial automation for ad-hoc data preparation tasks. They lack the "operationalization" capabilities of turning the preparation flows into a fully automated process, which can be built once, and reused afterwards.

Data preparation flows are iterative, and are 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).

K2View 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 warehouses in sync with your data sources, based on pre-defined data sync rules. Data changes can be ingested into your data stores in the delivery method of your choice: Extract, Transform, Load (ETL), data streaming, Change Data Capture (CDC), or messaging.



About K2View

K2View provides an operational data fabric dedicated to making every customer experience personalized and profitable.

The K2View platform continually ingests all customer data from all systems, enriches it with real-time insights, and transforms it into a patented Micro-Database™ – one for every customer. To maximize performance, scale, and security, every micro-DB is compressed and individually encrypted. It is then delivered in milliseconds to fuel quick, effective, and pleasing customer interactions. Global 2000 companies – including AT&T, Vodafone, Sky, and Hertz – deploy K2View in weeks to deliver outstanding multi-channel customer service, minimize churn, achieve hyper-segmentation, and assure data compliance.