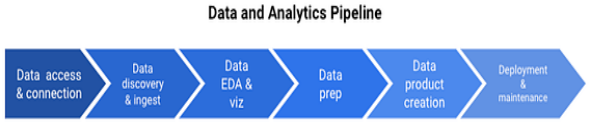
Moving From Good BI to Better BI to Even Better AI

Many organizations believe that they need to have all their data ducks lined up before they attempt AI analytics. They believe they need to have conquered traditional or business intelligence (BI) analytics first, including data catalogs, data lineage, master data management, big data, etc. before planning for AI. While this conventional thinking has merits, it results in high opportunity costs and carries risks. This article will debunk this common assumption and highlight how organizations – notably analysts and line of business managers – can establish capabilities for both traditional and AI analytics, leveraging common capabilities frameworks and tools that excel for both of them.

Data Pipelines Hold the Key

To move from data storage to BI insights (and, ultimately, business decisions), raw data needs to funnel through the data and analytics pipeline:



The process starts with getting access to data sources, establishing connections to them, and ensuring that you can discover various datasets. Then, you'll perform exploratory data analytics (EDA) and visualization to get a sense of the data. This is followed by data prep and data product creation (in the case of BI, data products include tables, charts, graphs, dashboards, etc.). Finally, if you›re planning to refresh and reuse the data products, you'll want to deploy them and the rest of the data pipeline and make sure they work in a production environment.

In an ideal world, the storage component would be fully built out and the data pipeline would be robust, so that confident analyses lead to rapid business decisions. With respect to storage, you'd want the data schemas and storage types, master data management, data quality monitoring, and the cloud/on-prem architecture strategy to be sorted.

For the data pipeline, you would have governance in place, full data lineage, robustness and scalability (regardless of the size of the data), high analytics throughput enabled via data wrangling and data product development, and the ability for decision makers to work with the data themselves (because they would, ideally, know it the best).

However, this is often not the case.

In reality, there's continuous adaptation on the storage side – new architecture strategies and data storage types continue to be developed, along with new schemas and new data types. For data pipelines, teams often ask themselves "What is the right pipeline that ingests the right datasets and uses the right transformations to create the right outputs?" This may stem from factors such as

limited visibility to data lineage, extensive use of unwieldy spreadsheets for data transformations, and bottlenecks and capacity constraints that lead to time-consuming iterations with end stakeholders (the data consumers). Unlike the scenario above, these issues may lead to slow, unsure decision making, driven by outdated or incorrect information and questions on the veracity of the data.

So, How Can We Navigate Those Problems?

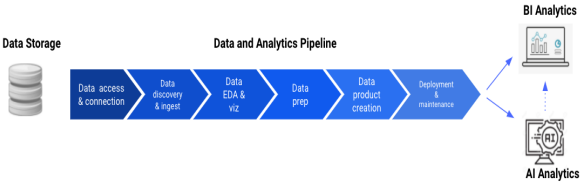
We believe that teams can enhance their data and analytics pipelines while populating and iterating on data storage (e.g., data lakes) – in other words, executing in parallel vs. sequentially – to significantly improve BI analytics and to make progress in AI analytics now. This is very doable provided that the data pipeline has the flexibility to connect to the data storage and to process the data types that will be stored. Additional characteristics of an enhanced data pipeline include:

- Higher efficiency (lower hours of effort) and more data products throughput
- Data lineage transparency and analytics reproducibility (reinforcing confidence in the data), reusability (to save time by leveraging what others have built), and robustness (so it's reliable and doesn't break in production)
- Self-service capabilities to eliminate staff bottlenecks and to speed decision making

If teams are able to accomplish this, their step from BI to AI is going to be smooth. Many data scientists spend roughly ٪٨٠ of their time/effort on getting the data, wrangling it, and preparing it for model development. These activities are the exact same activities as the first four in the data pipeline and, if the data pipeline is deployed and maintained in production, they represent five of the six activities.

In other words, by enhancing the BI data pipeline, you've almost built an AI data pipeline. The additional step to turn the BI data pipeline into an AI data pipeline is to create AI data products (i.e., machine learning models). And to transition from building BI data products to AI data products, staff can reuse the data assets and infrastructure they've already built and understand well, they can shorten the learning curve by leveraging an identical user experience in the data pipeline, and they can scale faster and more economically by doing both BI and AI projects in one environment.

As seen below, the same data pipeline for BI analytics can serve AI analytics:



data and analytics pipeline

Key Data Pipeline Functionalities to Incorporate

For the data pipelines and workflows themselves, make sure they:

- Are visual and transparent to all stakeholders (to enhance rapid iteration of projects)
- Have both no-code and code options available (to get non-technical experts productive)
- Enable reusable data assets and pipelines and recorded lineage for rollback and branching
- Offer easy deployment into production
- Promote rapid data discovery, connection, and ingestion
- Allow for quick statistical analysis of data fields and columns
- Enable the creation of multiple types of data products, including machine learning models, tables, charts, graphs, and dashboards

Once your data analytics pipeline is enhanced, you will not only be able to, but will want to, execute AI in parallel with BI. Why's that? Well, the first reason is economies of scope – BI and AI involve the same or much of the same data sources, common data understanding and interpretation, and the same data pipelines themselves. Further, teams can reduce the risk of rework in the future by including AI model needs now when building data sources and pipelines – things like missing or incomplete fields, poor quality of data fields, unlabeled data, etc. Finally, AI benefits realization takes time – there is a learning curve to developing and operationalizing AI models so the sooner you get started, the sooner you will see results.