The Data Scientist's Workbench

Logical Data Warehouse

The three elements below combine to create a logical data warehouse – the backbone of all analytical needs of an organization, as well as the technologies to best support them. This page introduces the data lake as a data source for enabling advanced analytics.

DATA WAREHOUSE

The data warehouse contains operationally strategic data to support key business processes. The data is typically utilized to support most of the business needs and is rigid in structure.

VIRTUALIZATION

Virtual data layers can be designed for ad-hoc agile analytical development with semantic layers typically covering department level analytic cohorts.

DATA LAKE

Economic storage for data to be loaded in native or near native format. Designed for schema-on-read to support the creative investigation of data scientists for advanced analytics initiatives.

How to Access the Data Lake:

By way of the data catalog, tools like PowerBI and Tableau integrate many different data sources to investigate the relationships of the raw data as it relates to the advanced analytics initiative. Although this stage is typically performed by those with great statistical and computational knowledge, easy to use tools like Tableau and PowerBI democratize insight discovery.

Conceptualizing the Data Lake:

The data lake is a technology designed to store data in its native or near native format. The benefit of storing data in this format is that there is no loss of information translating the data into a rigid structure. Data scientists easily investigate all possible relationships in the data to support the advanced analytics initiative. As the findings become more critical and strategic to the company, semantic layers can be created from the data lake to mature into virtualization layers and the data warehouse.

Benefits of the Data Lake:

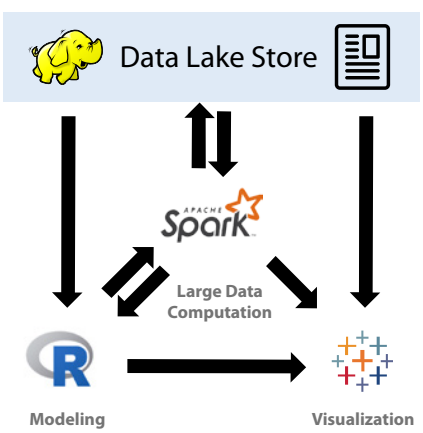
1. Economic storage for all types of data
2. No compromising structure requirements
3. Ability to sandbox and invite all data workers to interact
4. Cloud capabilities allow for all types of compute engines to access data
5. Positively disruptive findings can mature into virtualization and data warehouse sources

Technologies Empowering Research and Implementation Phase

Integrating tools used in both the research phase, and the implementation phase, automates the process of an insight found in the research phase moving into the implementation phase. The automation of the predictive pipeline through the coordination of these technologies enable businesses time to react to the insights quickly and automatically, creating economies at scale.

WHAT TO EXPECT IN THE RESEARCH PHASE

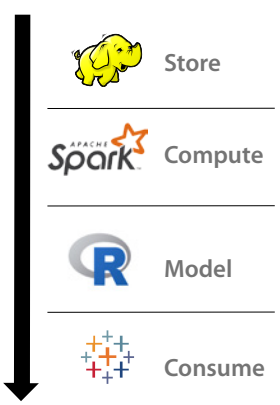
- Location for exploring data
- Rapid pre-processing of data for building models
- Platform to communicate findings
- Usage of widely known analytics tools



Technology mix enabling autonomy for discovery

WHAT TO EXPECT IN THE IMPLEMENTATION PHASE

- Effortless transition from lab to implementation
- Fast execution of processing and scoring
- Clear scheduling for an automation process
- Easily accessed by consumption layers



Technology mix for IT-enabled implementation and predictive pipeline

TECHNOLOGIES

Data Lake: Access and storage of raw data
Spark: Cluster computation for big data
R: Robust and prevalent machine learning tool
Tableau: Visual exploration and consumption

Technologies play a role in both the lab and implementation phase. The architecture works because of advances in tools that combine R and Spark. This affords the ability to standardize across tools, making scalability and enablement easier. These technologies are merely examples of a singular pipeline, so following iterations may look different. For example, we may need to use Hive to transform data before leveraging Spark.

Developers can significantly decrease time spent experimenting and modeling in Microsoft's robust R Server and by leveraging the computational speed of Spark.

CONTACT US TODAY

303.248.8321

www.neudesic.com