Outline of a Modern Data Warehouse

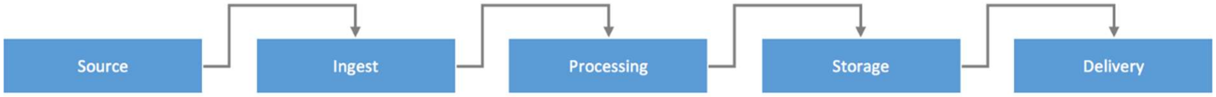
Data warehouses and data marts consist of data that has been transformed from sources into models that target specific areas of analysis and reporting. Typically, the end goal is to provide a single source of truth while allowing for advanced analytics--such as text mining and prediction (machine learning).

The classic definition of a data warehouse is one which embodies an Inmon or Kimball model built on a relational database management system (RDBMS). This is traditionally extended to include aggregation and analytics servers within the same environment. The concept of a "modern" data warehouse is big data focused and sits on a data storage system that allows for any type of data structure (images, text, random documents). The big data component relates to parallel processing of data with a Hadoop based server architecture on top of a distributed file system. This shift serves to address the performance issues that are often encountered when dealing with very large data sets in a RDBMS, and it consolidates advanced aggregation and analytics capabilities across a large ecosystem of open source tools. As a bonus, most of the principles learned or developed in traditional data warehousing paradigms--in years past--still hold true.

The desire to have all data centralized in order to capitalize on these data assets has driven organizations to develop data lakes. A data lake is a landing zone for all of an organization's raw/untransformed data. The modern data warehouse integrates the data lake, data marts, and specialized servers—Hadoop, MapReduce2, YARN, Hive, Spark, and R--to facilitate advanced analytics and reporting.

The following outline details the services and methods used to build a data analytics platform in Microsoft Azure.

The Data Pipeline



- **Source**
  - Data sources such as: SQL Server, streaming services, and flat files from partners and vendors.
- **Ingest**
  - Computation that involves copying or moving data to a central storage system and schematizing. Raw Schema.
- **Process**
  - Computation processes built by data engineers that transform raw data from the lake into specific data models that satisfy a particular analysis. Refined schema.
- **Store**
  - Data locations where ingested and processed data are stored either temporarily or permanently.
- **Deliver**
  - Presenting data to the end user.

Ingest

The target data store is a fault tolerant, secure, and infinitely scalable subsystem (Azure Data Lake Store) with a directory structure that will allow users to browse and query raw data files. These files are never modified or removed. This data store is the “Data Lake.”

Tools and Services: Visual Studio, Azure Data Lake Store, Azure Data Factory, SSIS, Powershell, HDInsight, Hive

Process

In addition to tying together, and making the data easier to analyze, this part of the pipeline serves to greatly enhance query performance as well. This point in the pipeline produces transformed copies of data from the data lake—collections of which may serve as data marts. A data mart is a collection of data that serves specific areas of the business.

Tools and Services: Visual Studio, Azure Data Lake Store, Azure Data Factory, HDInsight, Hive

Store

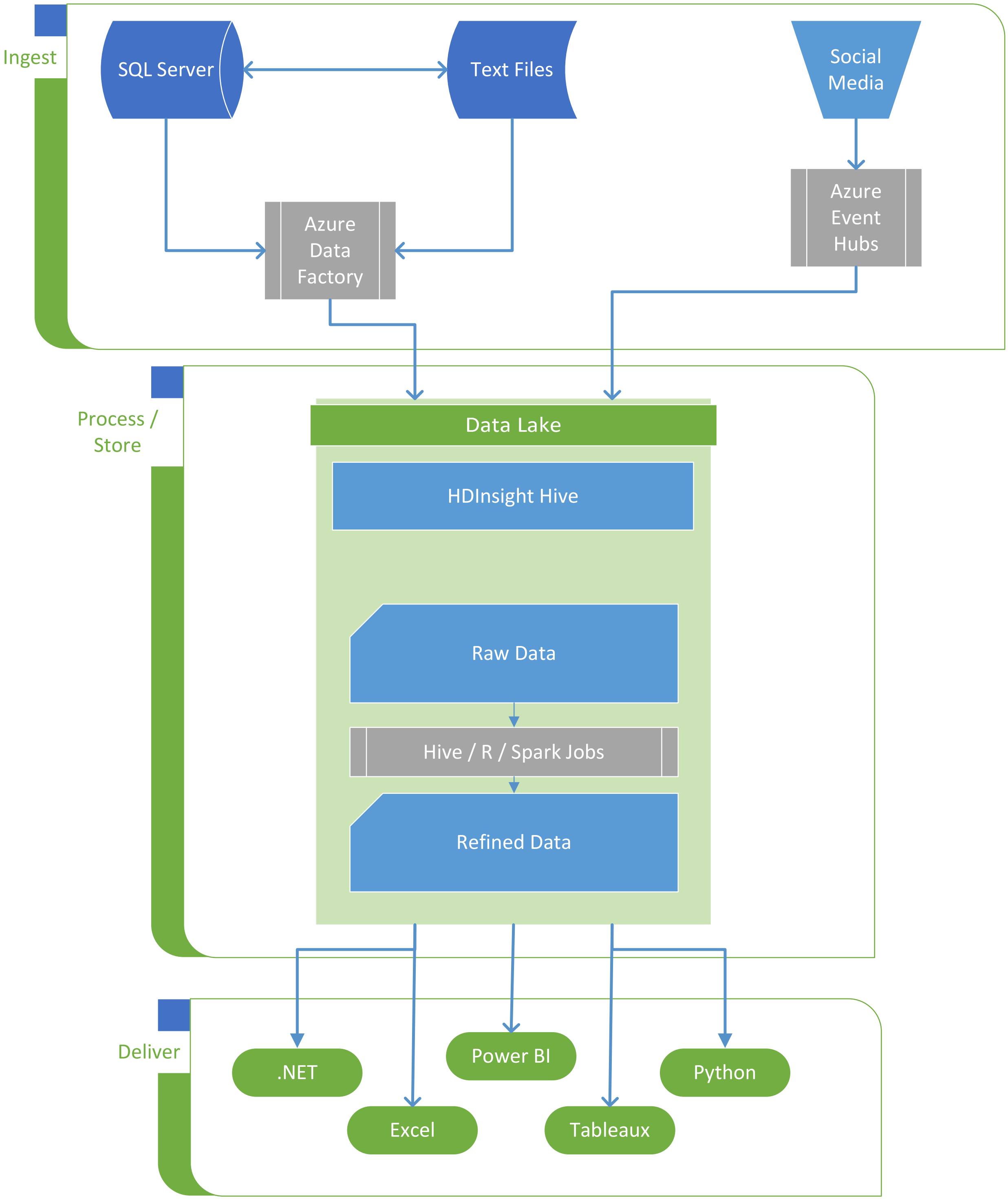
Azure Data Lake Store (ADLS) is the primary storage subsystem. Azure Blob storage may be utilized in some cases where ADLS is not yet compatible with tooling or services. Standard patterns of storing processed and unprocessed data on the distributed file system make it available to a wide range of server products that are easily provisioned in Azure on a permanent or as-needed basis. If there is a question as to capability the answer is usually “yes.”

Tools and Services: Visual Studio, Azure Data Lake Store, Azure Data Factory

Deliver

The beauty of the modern data warehouse is the full range of tooling that is available to end-users. Because the distributed file system and various methods of organizing and querying data across the Hadoop ecosystem is so standardized and integrated, the data is ready to be consumed by almost any tool or programming language available. Data engineers can build canned reports for less-sophisticated users but the real power is in self-service. The above pipeline components, when implemented correctly, create a powerful data analytics platform that can leverage the creative ability and talent of end-users.

Tools and Services: Excel, Tableaux, Power BI, R, Jupyter, C#, Scala, Java, Python, SQL, HQL, Ambari, U-SQL


Ingest
SQL Server
Text Files
Social Media
Azure Data Factory
Azure Event Hubs
Process / Store
Data Lake
HDInsight Hive
Raw Data
Hive / R / Spark Jobs
Refined Data
Deliver
.NET
Power BI
Python
Excel
Tableaux