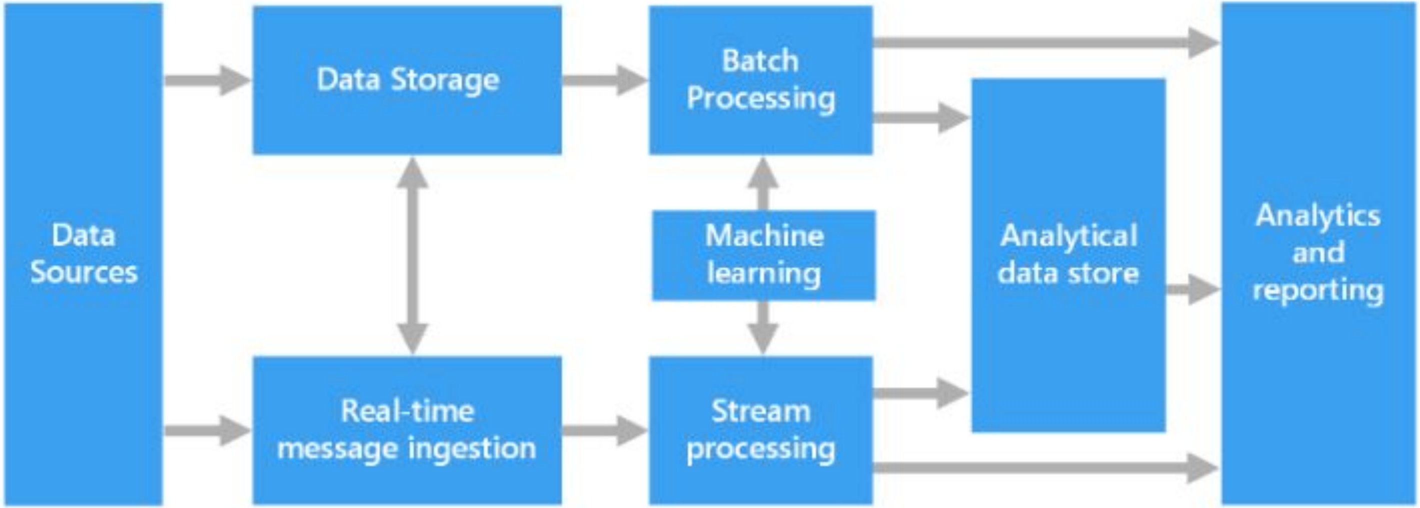
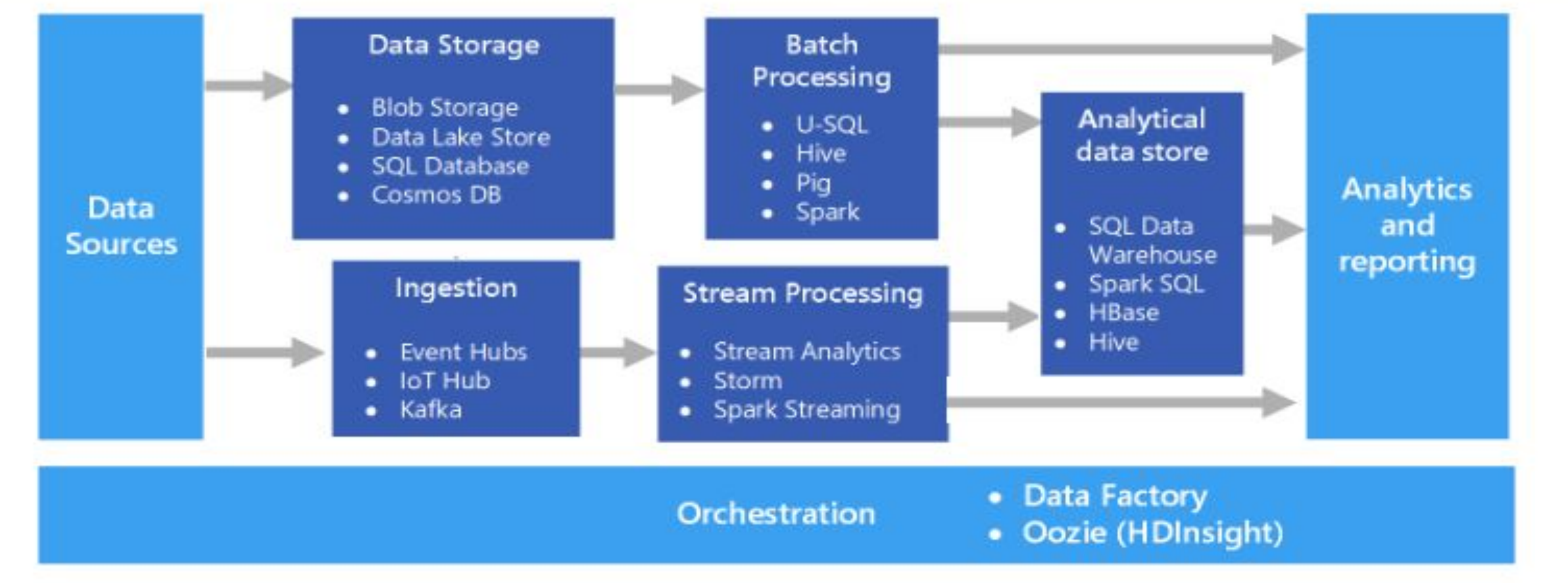
Architecture Design for Data Platform using Big data and Cloud

HLD Component



LLD Components



Description

This is lambda architecture platform for batch and real time data processing and designed approach is ELT based on Azure cloud and big data ecosystem.

Datasource: This architecture supports multiple source like RDBMS, flat files, cloud file system etc.

Ingestion Tools: There are multiple tools available to ingest data into cloud data lake for batch and realtime data processing. ADF, sqoop, jdbc/odbc, connectors are used to ingest data into azure data lake storage. Eventhub, kafka, IoT hub can be used as data ingestion tool to data lake in real time.

Transformation tools: Hive and spark is capable of tranforming data for batch piepeline. For real-time data transformation and ingestion stream analytics, Spark and storm can be used.

Cont..

Machine Learning Tools: Multiple libraries available for machine learning i.e for structured data Scikit-learn, for deep learning keras, tensor flow, pytorch. Azure cloud also provide azure ML service that used for development and tracking of project and choose best model among them. Similarly databricks also provide platform to develop and track experiment called MLFlow.

Analytical tool: Azure SQL DWH, Hive, Hbase or any cloud specific Nosql

Reporting tools: PowerBI, Tableau

Challenges

1. Multiple source comes with complexities like different schemas and some system specific data types
2. Firewall, networking and access issues to ingest data into cloud from on premise system
3. Authentication, authorization and overall security
4. Data encryption and masking at rest and in transit
5. Public cloud data security and governance
6. Multiple cloud provider and integration of different cloud platform
7. High availability and Disaster recovery
8. Scalability and cost optimization
9. Machine learning experiment and parameter tracking for different model, choose best model