## Interpreting the Architecture

### Query and Processing

<table>
<thead>
<tr>
<th>Sources</th>
<th>Ingestion and Transformation</th>
<th>Storage</th>
<th>Historical</th>
<th>Predictive</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generate relevant business and operational data</td>
<td>Extract data from operational systems (E)</td>
<td>Store data in a format accessible to query &amp; processing systems</td>
<td>Provide an interface for analysts and data scientists to derive insights (query)</td>
<td>Predict what will happen in the future</td>
<td>Present results of data analysis to internal and external users</td>
</tr>
<tr>
<td></td>
<td>Deliver to storage, aligning schemas between source and destination (L)</td>
<td>Optimize for low cost, scalability, and analytic workloads (e.g., column store)</td>
<td>Execute queries and data models against stored data, often using distributed compute (processing)</td>
<td>Build data-driven/ML applications</td>
<td>Embed data models into operational systems and applications</td>
</tr>
<tr>
<td></td>
<td>Transform data to a structure ready for analysis (T)</td>
<td>In some cases, provide additional data structures or guarantees</td>
<td>Describe what happened in the past (including very recent past)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Coordinate the flow of data and the execution of computations across the full lifecycle
- Ensure proper data quality, performance, and governance of all systems and datasets
Three Common Blueprints

Analytic Systems

1. Modern Business Intelligence

Multimodal Data Processing

2. Multimodal Data Processing

AI and ML

3. AI and ML

Operational Systems
1. Modern Business Intelligence Blueprint

**Query and Processing**

**Sources**
- OLTP Databases via CDC
- Applications/ERP (Oracle, Salesforce, Netsuite, ...)
- Event Collectors (Segment, Snowplow)
- Logs
- 3rd Party APIs (e.g., Stripe)
- File and Object Storage
- ...

**Ingestion and Transformation**
- Connectors (Fivetran, Stitch, Matillion)
- Data Modeling (dbt, LookML)
- Workflow Manager (Airflow, Dagster, Prefect)
- Spark Platform (Databricks, EMR)
- Python Libs (Pandas, Boto, Dask, Ray, ...)
- Batch Query Engine (Hive)
- Event Streaming (Confluent/Kafka, Pulsar, AWS Kinesis)
- Stream Processing (Databricks/Spark, Confluent/Kafka, Flink)

**Storage**
- Data Warehouse (Snowflake, BigQuery, Redshift)
- Data Lake (Databricks/Delta Lake, Iceberg, Hudi, Hive Acid)
- Parquet, ORC, Avro
- S3, GCS, ABS, HDFS

**Historical**
- Data Science Platform (Databricks, Domino, Sagemaker, DataRobot, Anaconda, ...)

**Predictive**
- Ad Hoc Query Engine (Presto, Dremio/Drill, Impala)
- Real-time Analytics (Imply/Druid, Altinity/Clickhouse, Rockset)

**Output**
- Dashboards (Looker, Superset, Mode, Tableau)
- Embedded Analytics (Sisense, Looker, cube.js)
- Augmented Analytics (Thoughtspot, Outlier, Anodot, Sisu)
- App Frameworks (Plotly Dash, Streamlit)
- Custom Apps
- ...

**Metadata Management**
- Collibra, Alation, Hive, Metastore, DataHub, ...

**Quality and Testing**
- Great Expectations

**Entitlements and Security**
- Privacera, Immuta

**Observability**
- Unravel, Accel Data, Fiddler
2. Multimodal Data Processing Blueprint

<table>
<thead>
<tr>
<th>Sources</th>
<th>Ingestion and Transformation</th>
<th>Storage</th>
<th>Historical</th>
<th>Predictive</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLTP Databases via CDC</td>
<td>Connectors (Fivetran, Stitch, Matillion)</td>
<td>Data Warehouse (Snowflake, BigQuery, Redshift)</td>
<td>Data Science Platform (Databricks, Domino, Sagemaker, Dataiku, DataRobot, Anaconda, ...)</td>
<td>Data Science and ML Libraries (Pandas, Numpy, R, Dask, Ray, Spark, ... Scikit-learn, Pytorch, TensorFlow, Spark ML, XGBoost, ...)</td>
<td>Dashboards (Looker, Superset, Mode, Tableau)</td>
</tr>
<tr>
<td>Applications/ERP (Oracle, Salesforce, Netsuite, ...)</td>
<td>Data Modeling (dbt, LookML)</td>
<td></td>
<td></td>
<td></td>
<td>Embedded Analytics (Sisense, Looker, cube.js)</td>
</tr>
<tr>
<td>Event Collectors (Segment, Snowplow)</td>
<td>Workflow Manager (Airflow, Dagster, Prefect)</td>
<td></td>
<td></td>
<td></td>
<td>Augmented Analytics (Thoughtspot, Outlier, Anodot, Sisu)</td>
</tr>
<tr>
<td>3rd Party APIs (e.g., Stripe)</td>
<td>Python Libs (Pandas, Boto, Dask, Ray, ...)</td>
<td>Databricks/Delta Lake, Iceberg, Hudi, Hive Acid</td>
<td></td>
<td></td>
<td>Custom Apps</td>
</tr>
<tr>
<td>File and Object Storage</td>
<td>Batch Query Engine (Hive)</td>
<td>Parquet, ORC, Avro</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>Event Streaming (Confluent/Kafka, Pulsar, AWS Kinesis)</td>
<td>S3, GCS, ABS, HDFS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stream Processing (Databricks/Spark, Confluent/Kafka, Flink)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Metadata Management (Colibra, Alation, Hive, Metastore, DataHub, ...) | Quality and Testing (Great Expectations) | Entitlements and Security (Privacera, Immuta) | Observability (Unravel, Accel Data, Fiddler) |
3. AI and ML Blueprint

Data Sources
(Data lake + data warehouse + streaming engine)

Dataflow Automation
(Airflow, Pachyderm, Elementl, Prefect, Tecton, Kubeflow)

Query Engines
(Presto, Hive)

Data Science Libraries
(Spark, Pandas, NumPy, Dask)

Data Science Platform
(Jupyter, Databricks, Domino, Sagemaker, DataRobot, H2O, Colab, Deepnote, Noteable)

Experiment Tracking
(Weights and Biases, Comet, MLflow)

Visualization
(Tensorboard, Fiddler)

Model Tuning
(Sigopt, hyperopt, Ray Tune)

ML Framework
(Scikit-learn, XGBoost, MLlib)

DL Framework
(TensorFlow, Keras, PyTorch, H2O, Hugging Face)

RL Libraries
(Gym, Dopamine, RLlib, Coach)

Distributed Processing
(Spark, Ray, Dask, Distributed TF, Kubeflow, Horovod)

Compiler
(TVM)

Model Registry
(Algorithmia, MLflow, Sagemaker)

Batch Predictor
(Spark)

Online Model Server
(TF Serving, Ray Serve, Seldon)

Model Monitoring
(Fiddler, Arthur, Arize)

Clients

Data Labeling
(Labelbox, Snorkel, Scaie, Sagemaker)