# A Unified Data Infrastructure Architecture

## 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)
- S3, GCS, ABS, HDFS

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

## Predictive
- Data Science Platform (Databricks, Domino, Sagemaker, Dataiku, DataRobot, Anaconda, ...)

## 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

## Additional Components
- Metadata Management (Collibra, Alation, Hive, Metastore, DataHub, ...)
- Quality and Testing (Great Expectations)
- Entitlements and Security (Privacera, Immuta)
- Observability (Unravel, Accel Data, Fiddler)
Interpreting the Architecture

<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>Deliver to storage, aligning schemas between source and destination (L)</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>Transform data to a structure ready for analysis (T)</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

1. Modern Business Intelligence
2. Multimodal Data Processing
3. AI and ML
2. Multimodal Data Processing Blueprint

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

Storage
- Data Warehouse (Snowflake, BigQuery, Redshift)
- Data Lake
  - 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)

Historical
- Data Science Platform (Databricks, Domino, SageMaker, Dataiku, DataRobot, Anaconda, ...)
- Data Science and ML Libraries (Pandas, NumPy, R, Dask, Ray, Spark, ... Scikit-learn, PyTorch, TensorFlow, Spark ML, XGBoost, ...)

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
- ...
### 3. AI and ML Blueprint

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

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

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

#### Query Engines
- (Presto, Hive)

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

#### Data Labeling
- (Labelbox, Snorkel, Scale, Sagemaker)

#### Experiment Tracking
- (Weights and Biases, Comet, MLflow)

#### Visualization
- (Tensorboard, Fiddler)

#### DL Framework
- (TensorFlow, Keras, PyTorch, H2O)

#### Feature Store
- (Tecton)

#### Feature Server
- (Tecton, Cassandra)

#### Feature Server
- (Tecton, Cassandra)

#### Model Registry
- (Algorithmia, MLflow, Sagemaker)

#### Compiler
- (TVM)

#### Model Monitoring
- (Fiddler,Arthur, Arize)

#### Model Inference
- (Spark)

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

#### Batch Predictor
- (Spark)

#### Model Tuning
- (Sigopt, hyperopt, Ray Tune)

#### RL Libraries
- (Gym, Dopamine, RLlib, Coach)

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