Data warehousing

OLAP : On-Line Analytical Processing

A. Espinosa 24/02/25

Transaction processing

Operations were commercial transactions

- product purchase
- placing an order with a supplier
- paying an employee’s salary

Applications get more complex

Transaction definition becomes more generic:

- Comment on blog post
- Action in a videogame
- Local NFC operation

OnLine Transaction Processing (OLTP)

- Interactive applications (on line)
- Applications look up for a small number of records
- Records are inserted or updated

- Many examples:
  - Favorite songs in Spotify
  - Grades in academic report
  - NCBI gene search by gene name
  - last 10 operations of my bank account

Data analytics questions

- What was the total revenue of each store in january?
- How many more products were sold in last year black friday?
- How many people have died because of COVID since 2019?

- Business analysts generate reports
- Make better decisions: business intelligence

OnLine Analytic Processing (OLAP)

Different access patterns than traditional OLTP

- Scan over large number of records
- Read few columns per record
- Calculate some aggregate statistics
- Not returning raw data to the user

OLTP vs OLAP

Property	Transaction processing systems OLTP	Analytic systems OLAP
Main read pattern	Small number of records per query, fetched by key	Aggregate over large number of records
Main write pattern	random-access, low latency writes from user input	Bulk import (ETL) or event stream
Primarily used by	End user/customer, via client application (web)	Internal analyst, for decision support
What data represents	Latest state of data	History of events that happened over time
Dataset size	Gigabytes to Terabytes	Terabytes to Petabytes

Transactional vs analytical data stores

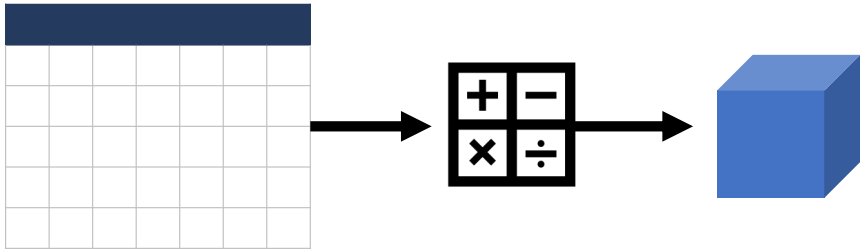
Online Transactional Processing (OLTP)

Customer		
CustomerID	CustomerName	CustomerPhone

Orders		
OrderID	CustomerID	OrderDate

Data is stored one transaction at a time

Online Analytical Processing (OLAP)

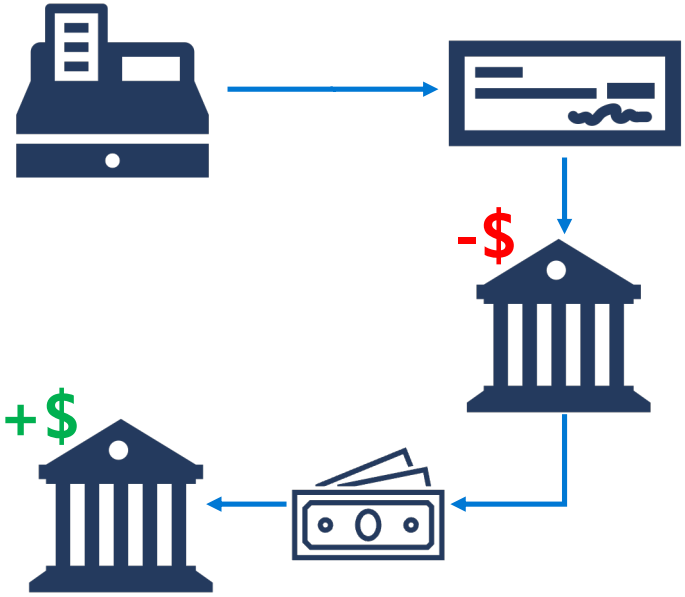


Data is periodically loaded, aggregated and stored in a cube

Transactional workloads

Transactional data is information that tracks the interactions related to an organization's activities.

- **Atomicity** – each transaction is treated as a single unit, which success completely or fails completely.
- **Consistency** – transactions can only take the data in the database from one valid state to another.
- **Isolation** – concurrent execution of transactions leave the database in the same state.
- **Durability –** once a transaction has been committed, it will remain committed.



Analytical Workloads

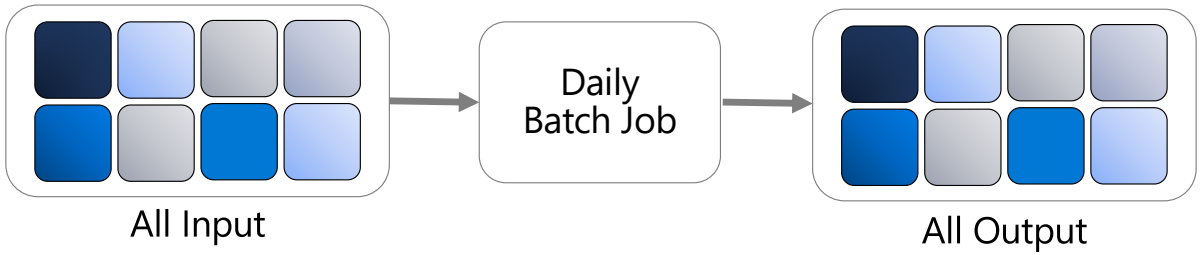
Analytical workloads are used for data analysis and decision making.

- Summaries
- Trends
- Business information

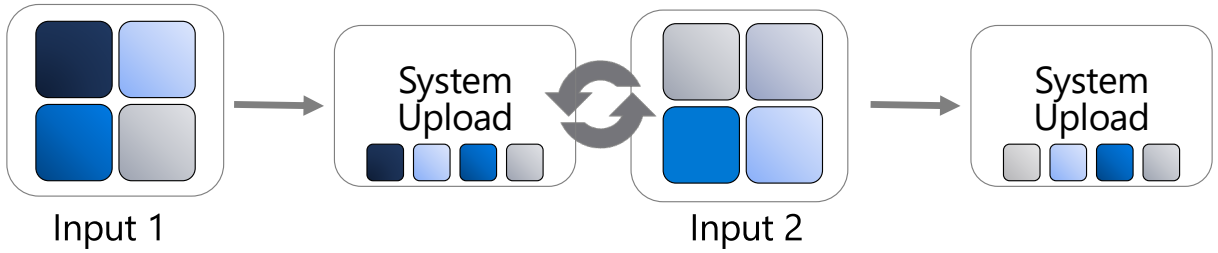
Data Processing

Data processing is the conversion of raw data to meaningful information through a process.

Batch Processing: data elements are collected into a group. The whole group is then processed at a future time as a batch



Stream Processing: each new piece of data is processed when it arrives.



Data warehouse

- Less use of OLTP systems for analytics
- Better: run analytics on a separate database
- Data warehouse
  - NOT affect main business database
  - Dump data to a new analytics data base for analysis
  - Ready to receive data intensive questions

Data system roles

Administrator, engineer, analyst

Job roles in data processing

• Database Administrator

Database Management

Implements Data Security

Backups

User Access

Monitors performance

• Data Engineer

- Data Pipelines and processes
- Data Ingestion storage
- Prepare data for Analytics
- Prepare data for analytical processing

• Data Analyst

- Provides insights into the data
- Visual Reporting
- Modeling Data for Analysis
- Combines data for visualization and analysis

Common tools – Database administrator

• Azure Data Studio	• SQL Server Management Studio	• Azure Portal/CLI
Graphical interface for managing on-premises and cloud-based data services	• Graphical interface for managing on-premises and cloud-based data services	• Tools for management and provisioning of Azure Data Services
	• Comprehensive Database Administration tool	• Manual and automation of scripts using Azure Resource Manager or Command Line Interface scripting

Common tools – Data engineering

• Azure Synapse Studio

Azure Portal integrated to manage Azure Synapse

Data Ingestion (Azure Data Factory)

Management of Azure Synapse assets (SQL Pools/Spark Pool)

• SQL Server Management Studio

- Graphical interface for managing on-premises and cloud-based data services
- Comprehensive Database Administration tool

• Azure Portal/CLI

- Tools for **management and provisioning of resources**
- **Manual and automation of scripts** using Resource Manager or Command Line Interface scripting

Common tools – Data analyst

• Power BI Desktop

Data Visualization tool

Model and Visualize Data

Management of Azure Synapse assets (SQL Pools/Spark Pool)

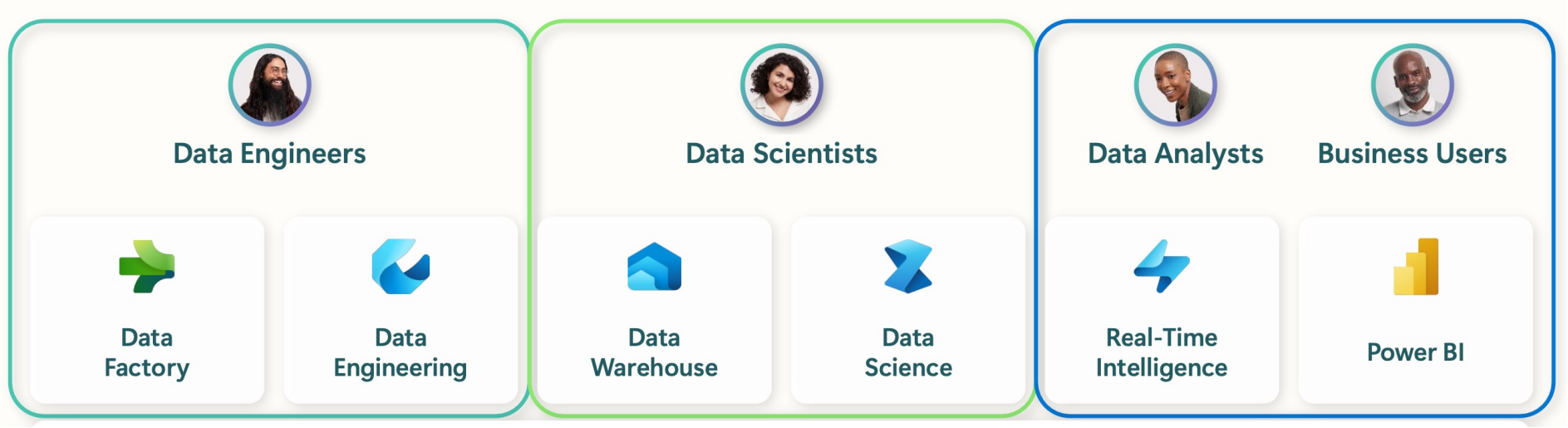
• Power BI Portal/ Power BI Service

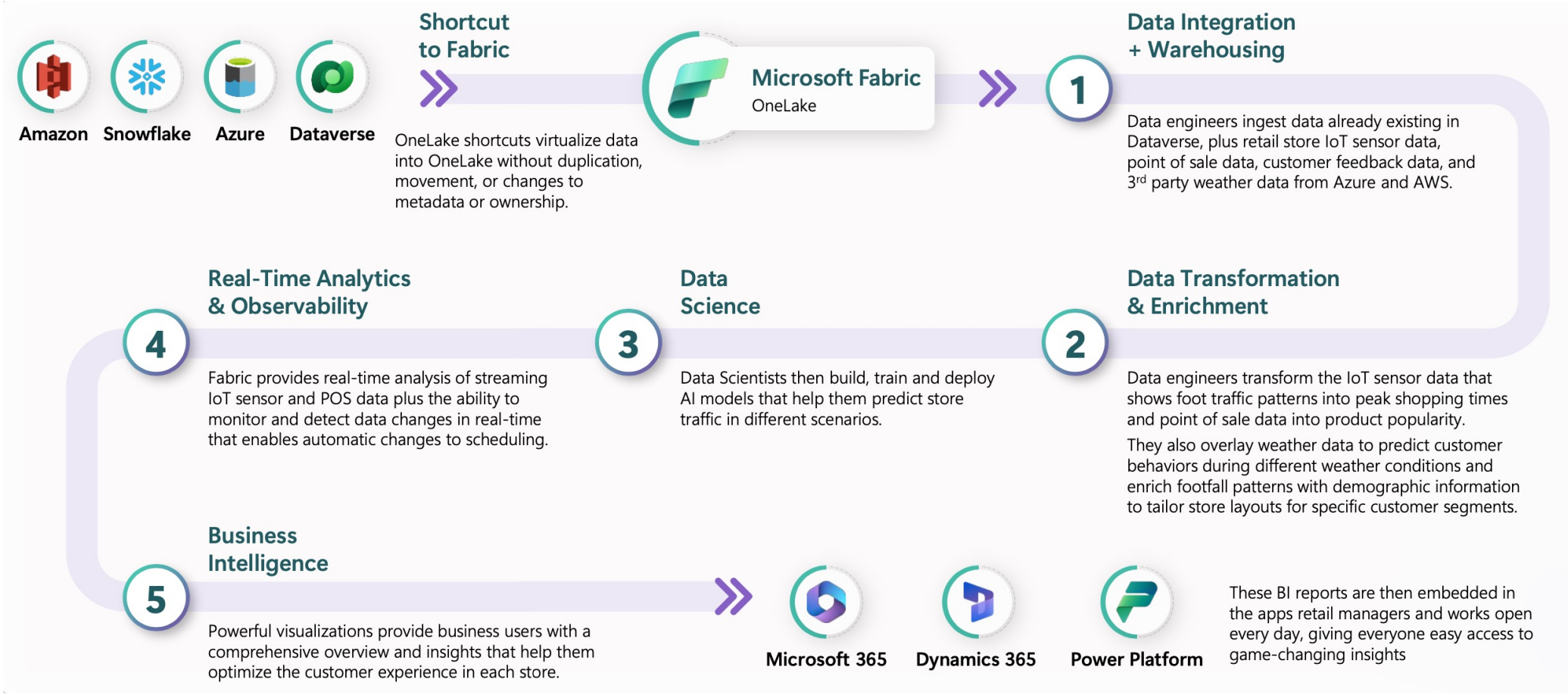
- Authoring and management of Power BI reports
- **Authoring of Power BI dashboards**
- Share Reports/Datasets

• Power BI Report Builder

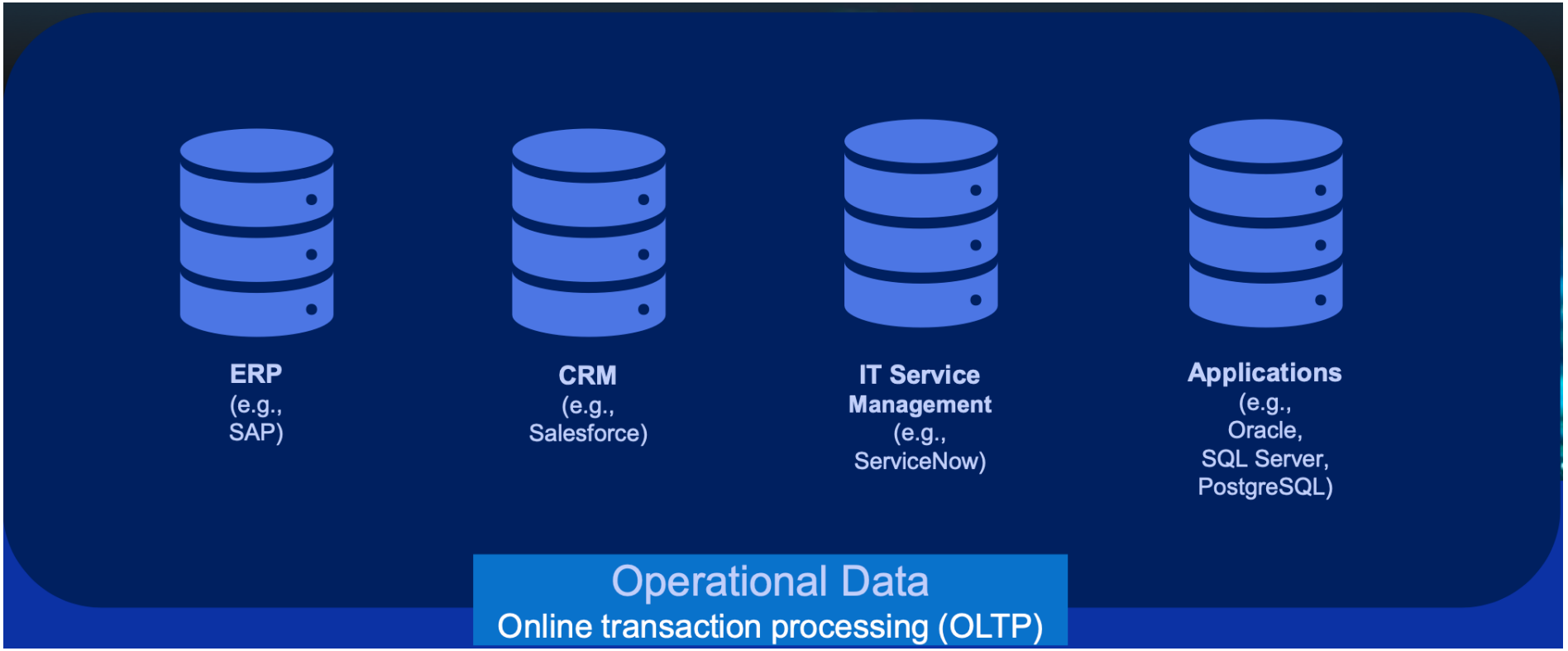
- Data Visualization tool for paginated reports
- **Model and Visualize paginated reports**

Roles in practic



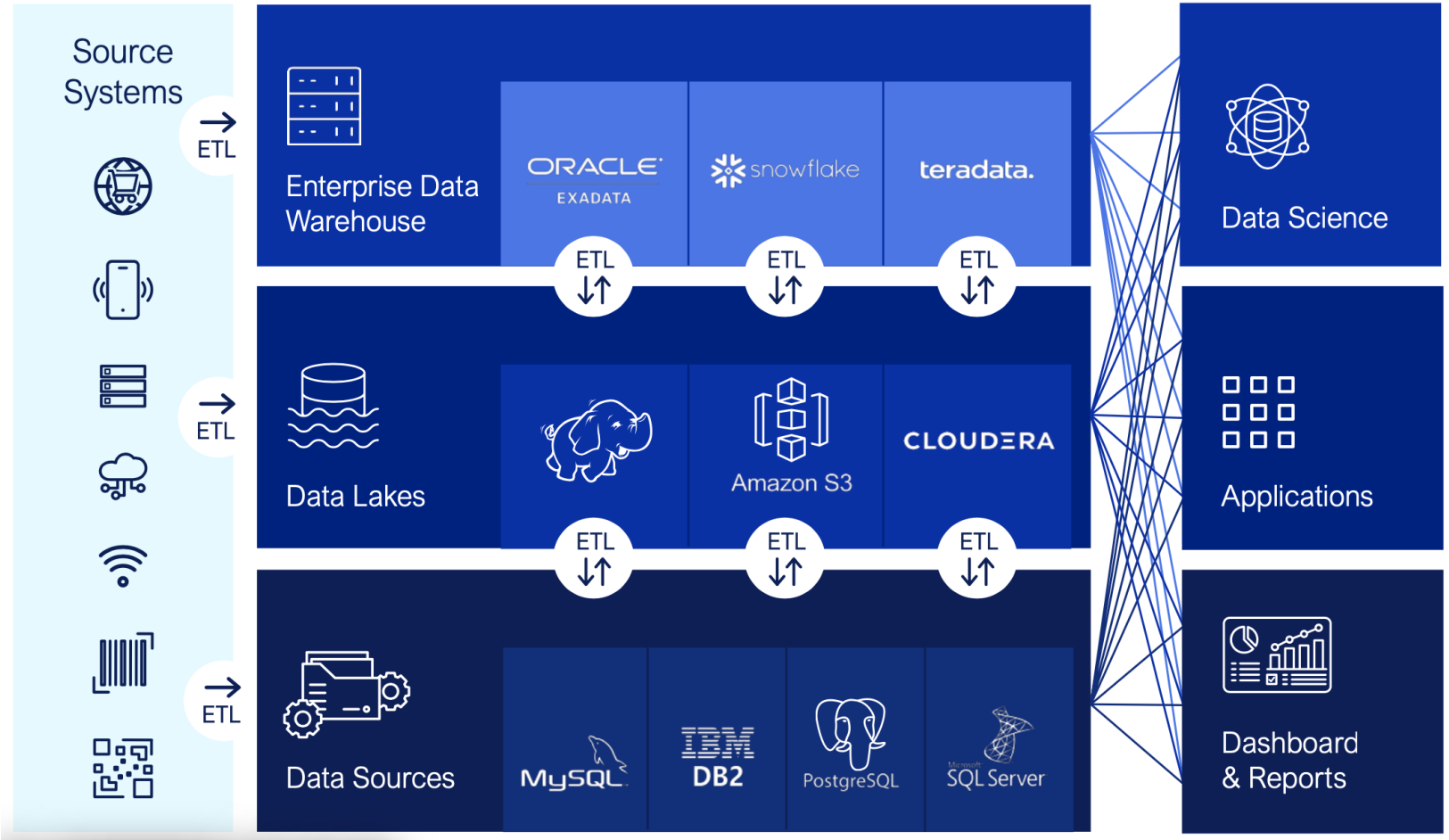

Amazon
Snowflake
Azure
Dataverse
Shortcut to Fabric
OneLake shortcuts virtualize data into OneLake without duplication, movement, or changes to metadata or ownership.
Microsoft Fabric
OneLake
1
Data Integration + Warehousing
Data engineers ingest data already existing in Dataverse, plus retail store IoT sensor data, point of sale data, customer feedback data, and 3rd party weather data from Azure and AWS.
2
Data Transformation & Enrichment
Data engineers transform the IoT sensor data that shows foot traffic patterns into peak shopping times and point of sale data into product popularity.
They also overlay weather data to predict customer behaviors during different weather conditions and enrich footfall patterns with demographic information to tailor store layouts for specific customer segments.
3
Data Science
Data Scientists then build, train and deploy AI models that help them predict store traffic in different scenarios.
4
Real-Time Analytics & Observability
Fabric provides real-time analysis of streaming IoT sensor and POS data plus the ability to monitor and detect data changes in real-time that enables automatic changes to scheduling.
5
Business Intelligence
Powerful visualizations provide business users with a comprehensive overview and insights that help them optimize the customer experience in each store.
Microsoft 365
Dynamics 365
Power Platform
These BI reports are then embedded in the apps retail managers and works open every day, giving everyone easy access to game-changing insights

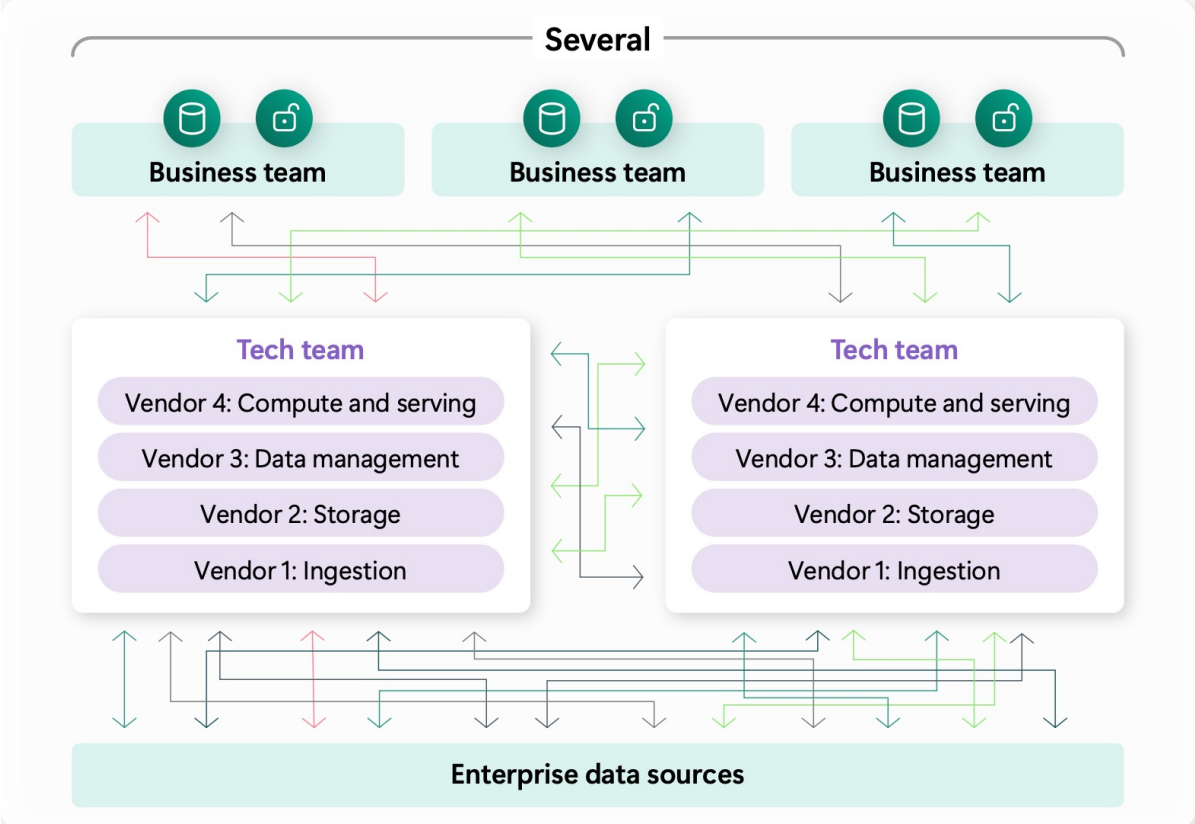
Sources of operational data



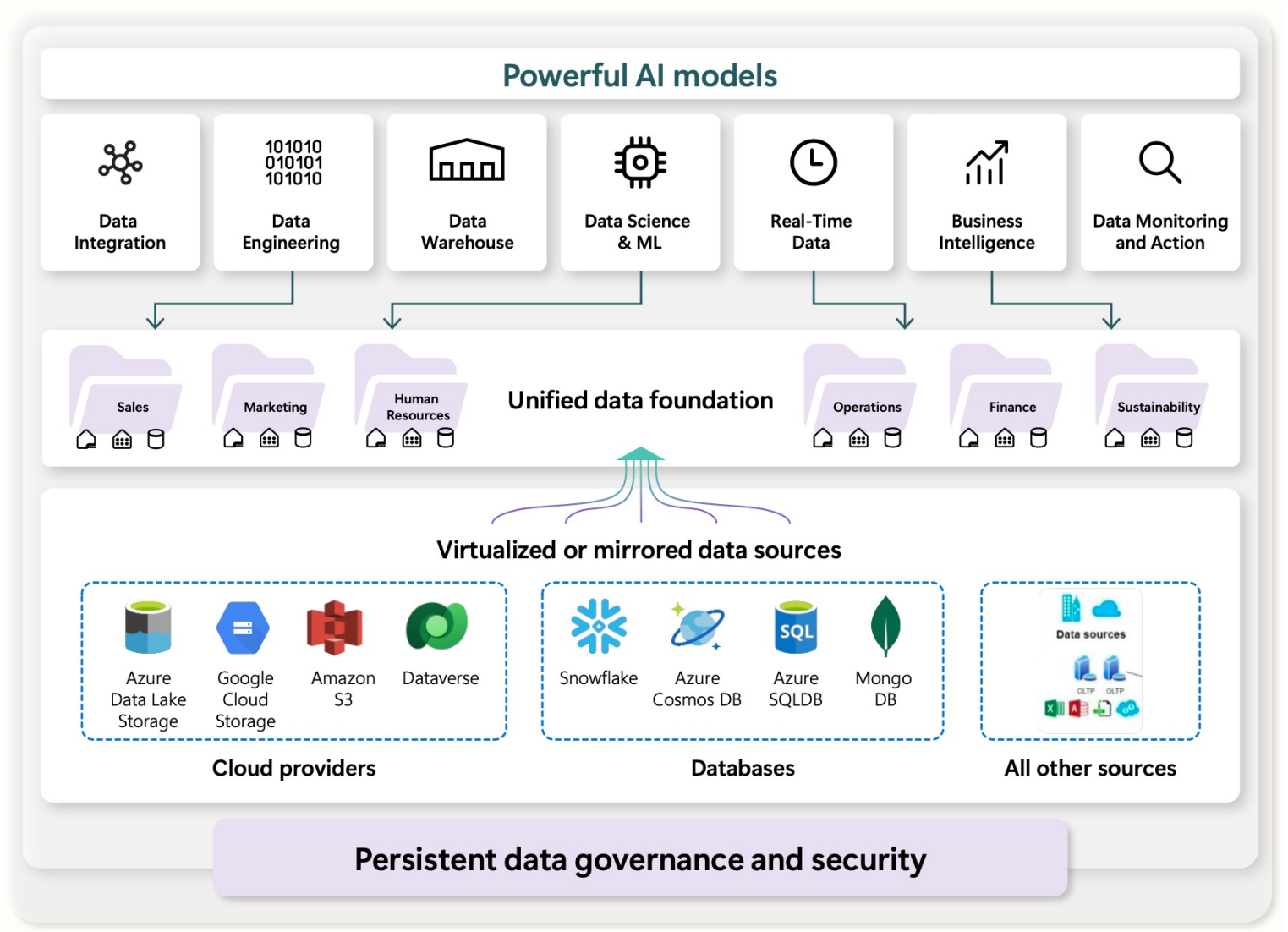
Modern tools for customer/provider mgmt

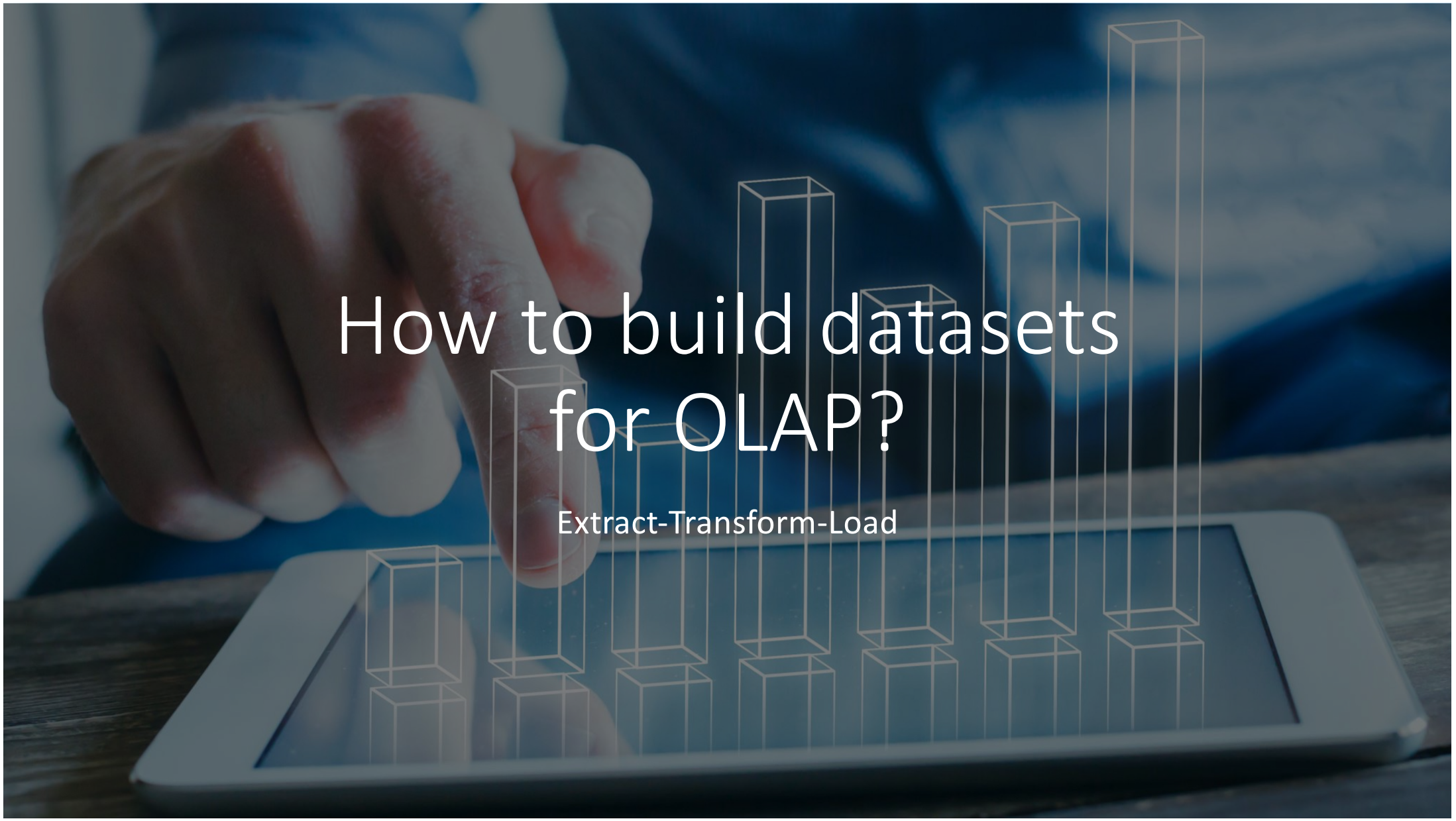
- **ERP: Enterprise Resource Planning**
  - **SAP, Microsoft Dynamics**
  - https://www.microsoft.com/es-es/dynamics-365#productdemos
  - https://www.sap.com/spain/products/erp/what-is-sap-erp.html
- **CRM: Customer Relationship Management**
  - **Salesforce, Odoo**
  - https://www.salesforce.com/mx/crm/
  - https://www.salesforce.com/es/customer-success-stories
  - https://www.odoo.com/es_ES/app/crm
- **Databases:**
  - E-commerce analytics
  - E-mail campaigns
  - Logistics data


Source Systems
ETL
ETL
ETL
Enterprise Data Warehouse
ORACLE EXADATA
snowflake
teradata.
Data Lakes
Amazon S3
CLOUDERA
Data Sources
MySQL
IBM DB2
PostgreSQL
Microsoft SQL Server
ETL
ETL
ETL
ETL
ETL
ETL
Data Science
Applications
Dashboard & Reports

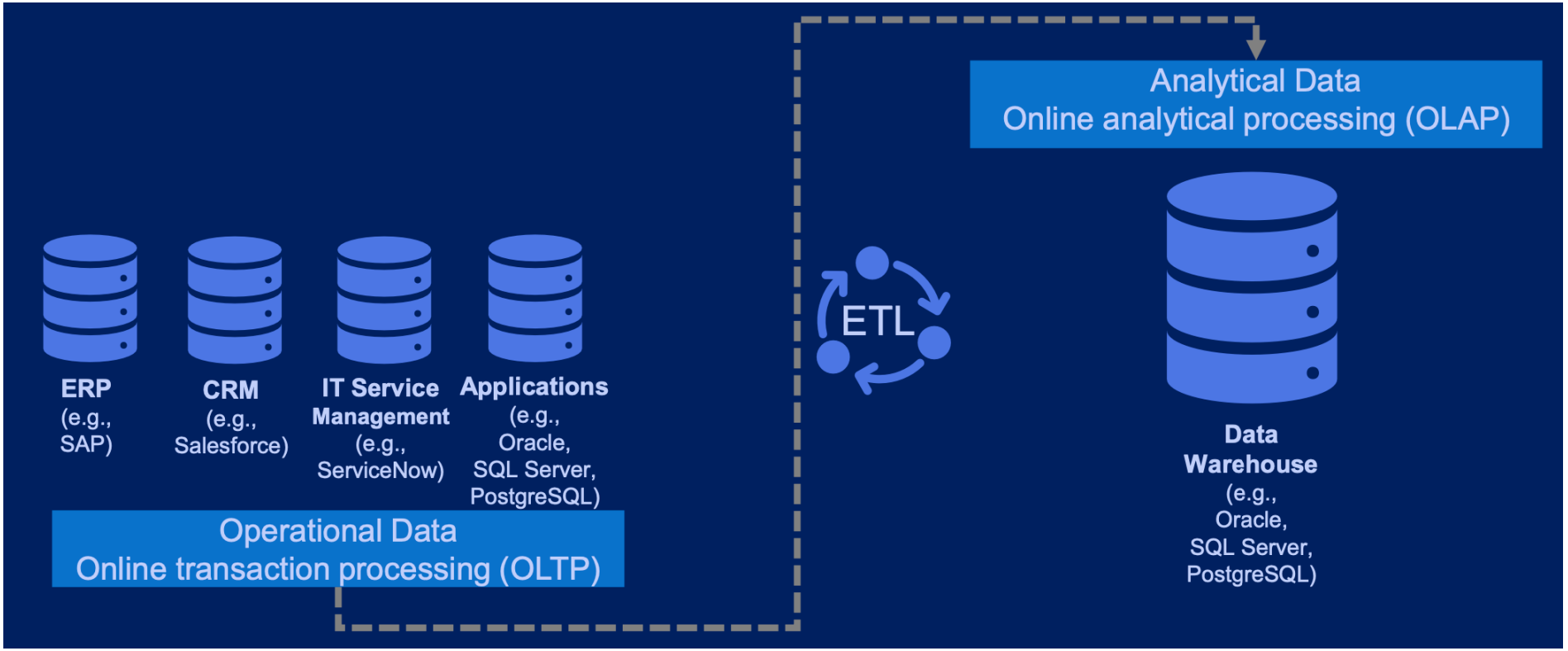


1. Data copies and infrastructure inefficiencies
2. Limited interoperability between vendor services
3. Data exposure risks


Powerful AI models
Data Integration
Data Engineering
Data Warehouse
Data Science & ML
Real-Time Data
Business Intelligence
Data Monitoring and Action
Sales
Marketing
Human Resources
Unified data foundation
Operations
Finance
Sustainability
Virtualized or mirrored data sources
Azure Data Lake Storage
Google Cloud Storage
Amazon S3
Dataverse
Snowflake
Azure Cosmos DB
Azure SQLDB
Mongo DB
Data sources
OLTP
OLTP
Cloud providers
Databases
All other sources
Persistent data governance and security


How to build datasets
for OLAP?
Extract-Transform-Load

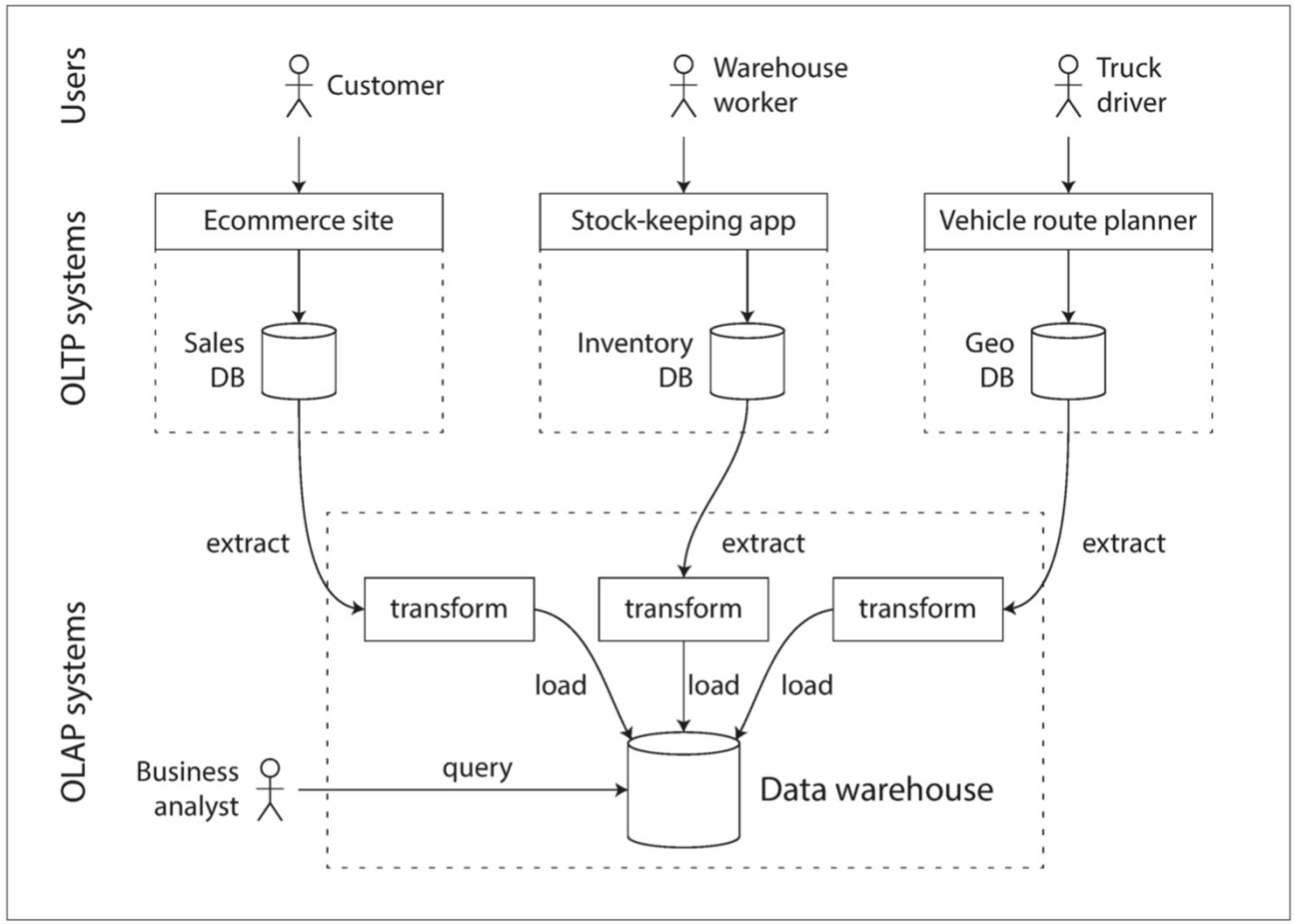
data integration for analytics



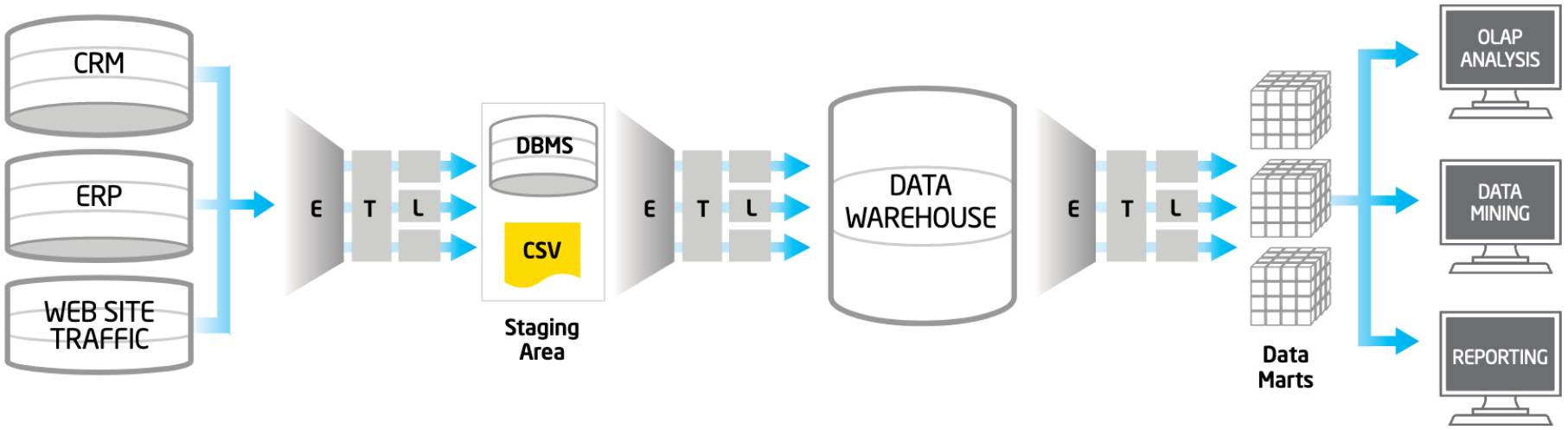
Extract-Transform-Load (ETL)

- How to build large datasets for analytics?
- Read-only copy of data in different OLTP systems in a company
- Main operations to apply to datasets:
  - Extract input from OLTP databases
  - Transform data into an analysis-friendly schema
  - Clean-up and filter our errors
  - Load data into the warehouse

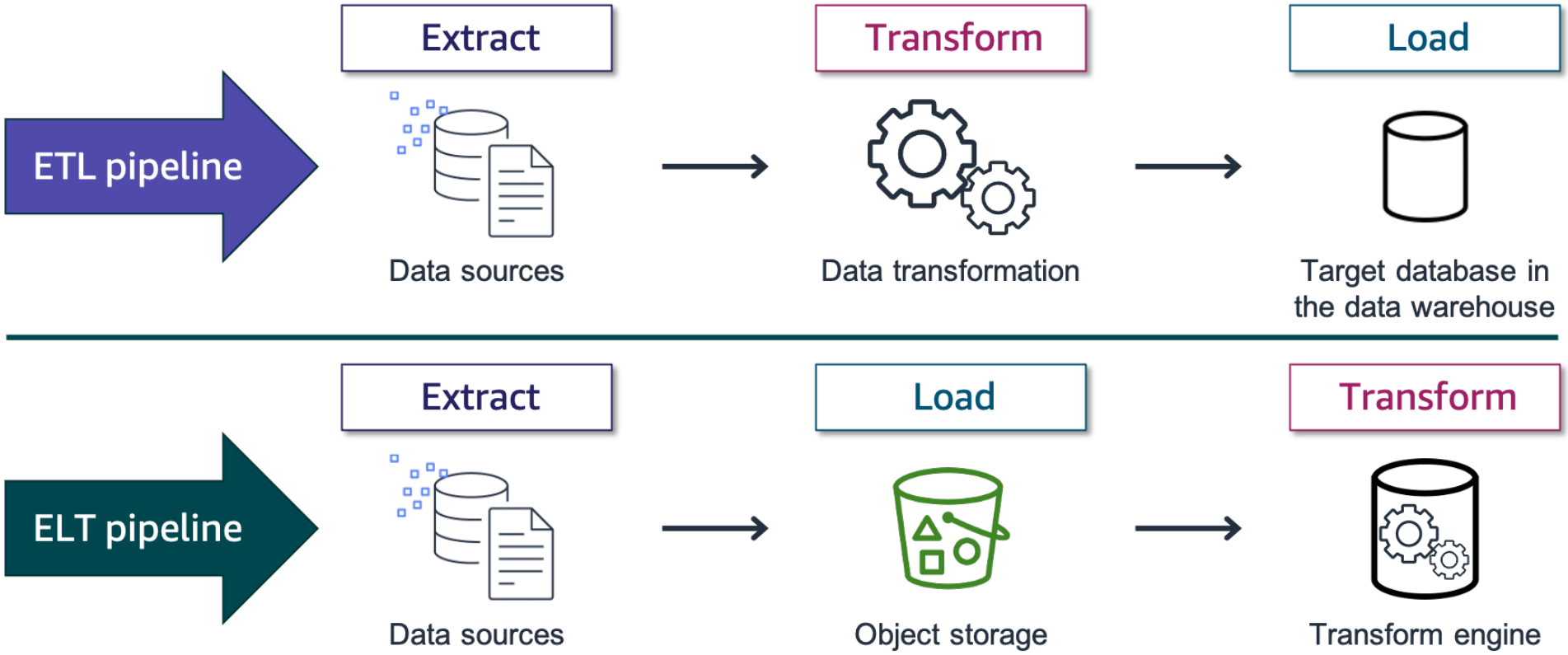
Simplified ETL



Data warehouse schema

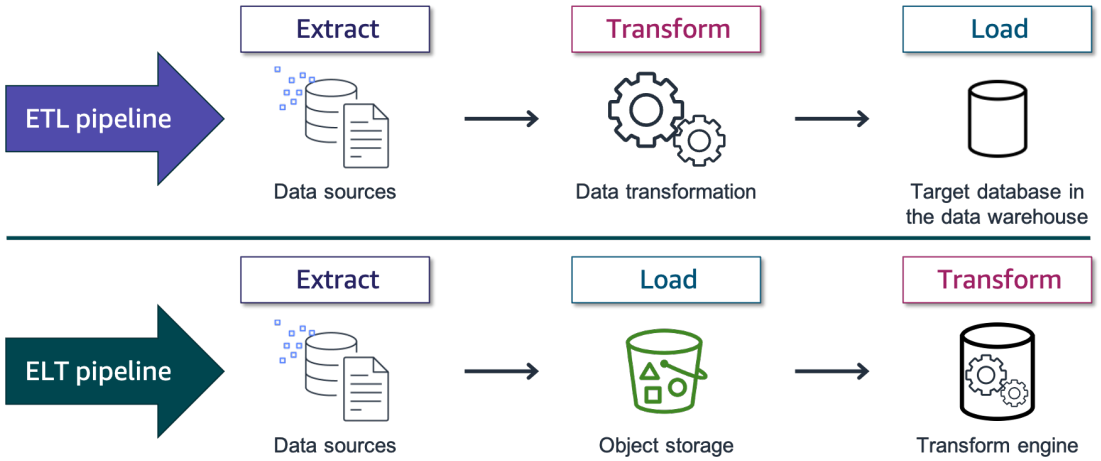


Comparing storage in ETL and ELT pipelines



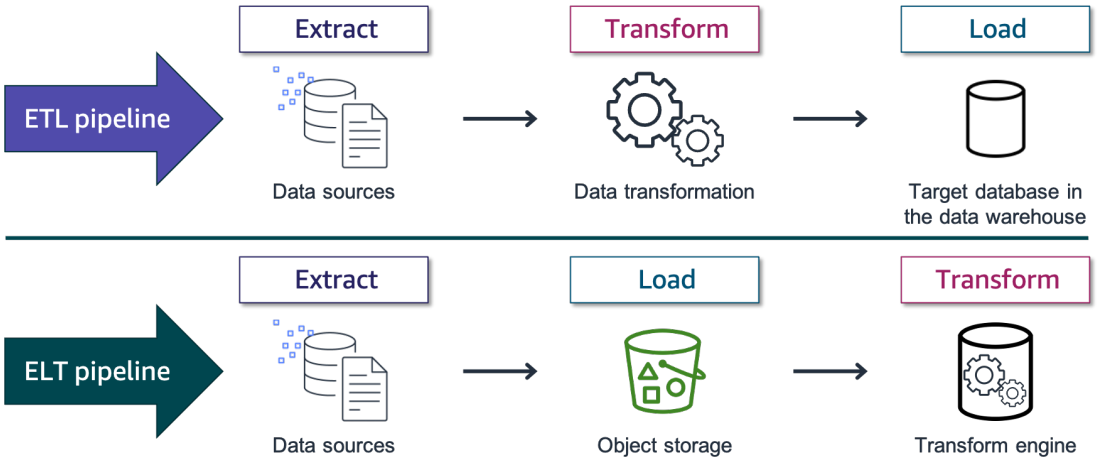
Traditional ETL pipeline: transform data first to save storage space

- Data is extracted from its source and transformed into a structured format.
- This transformed data is then loaded into structured storage, such as a data warehouse to perform additional transformations and processing if needed.
- Transformation is completed *before* it is stored in a data warehouse.

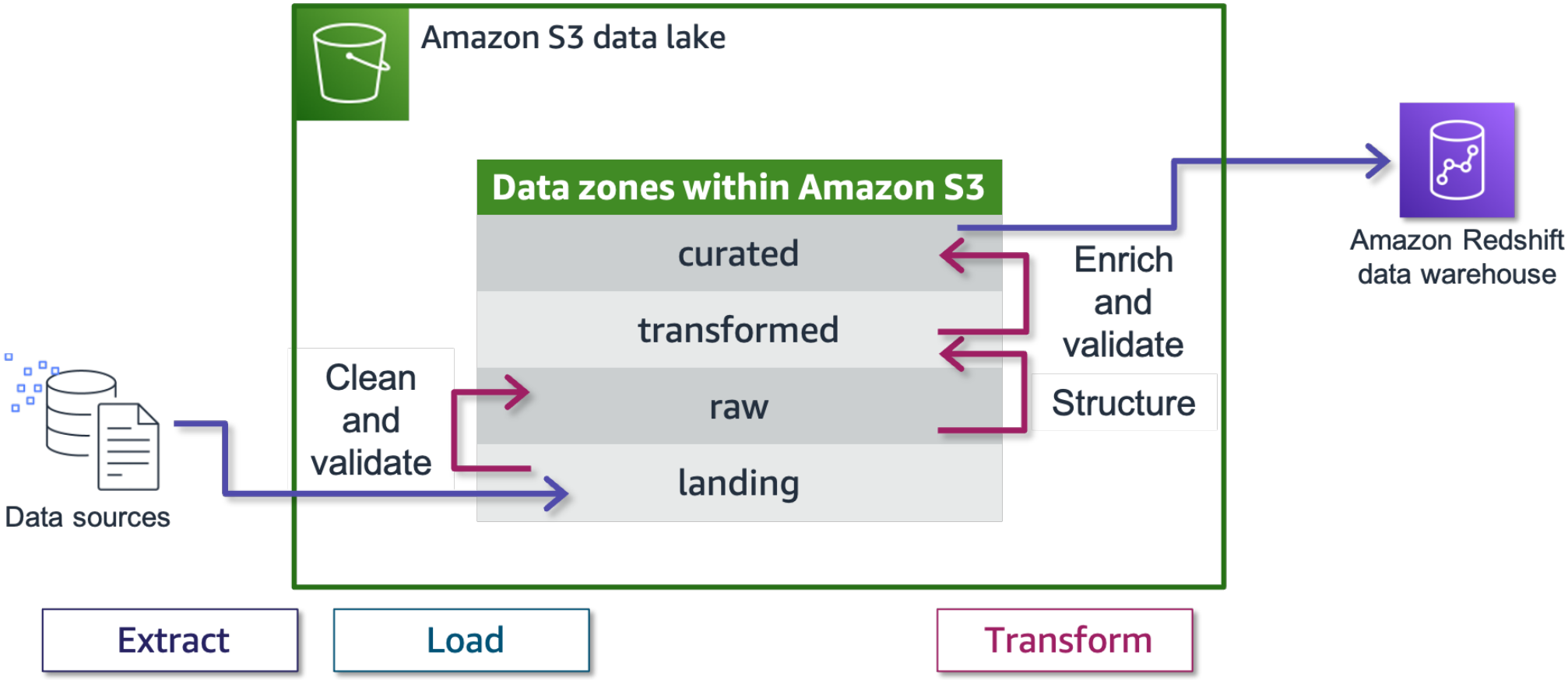


Modern ELT pipeline: Load everything then transform

- Data is extracted from its source and cleaned just enough to be stored.
- Data transformation engine is built into the data warehouse for relational and SQL workloads accesses all the data.
- This pattern needs resources: optimized and scalable data storage and compute power of massively parallel processing (MPP) architecture.



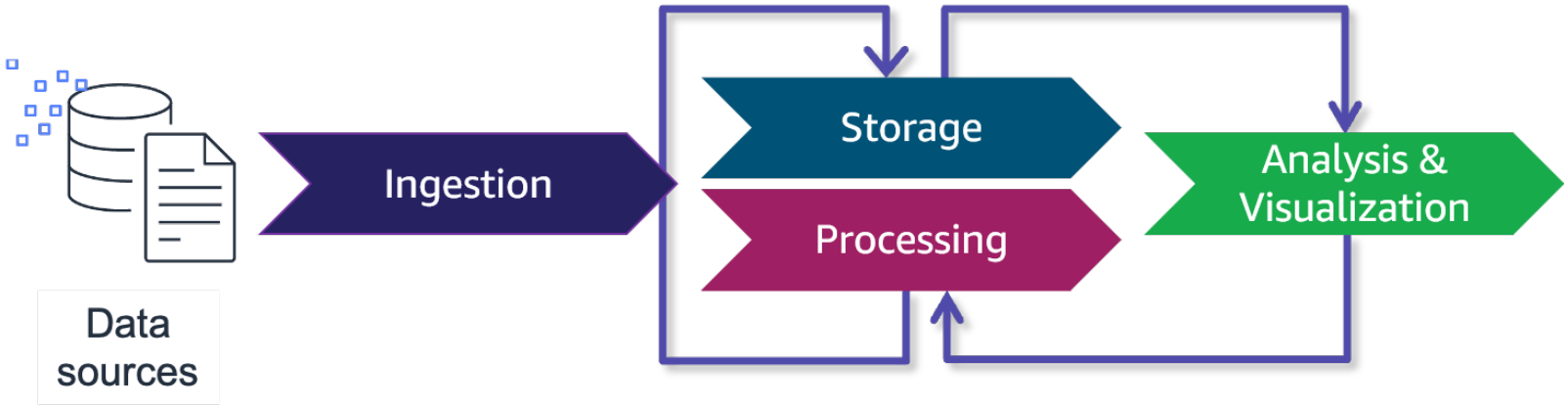
Example: Storage in support of an ELT pipeline



ELT pipeline architecture example

- Data sources are ingested and immediately sent to an Amazon S3 landing zone within a data lake.
- This data is cleaned and validated, and then transferred within the data lake to a raw data zone.
- As data is further processed, it is loaded, transformed, and moved into the appropriate S3 zones until it is ready for implementation in a data warehouse setting or for use in an analytics engine.
- While the ELT pipeline simplifies the architecture, the burden of the transformation workload is placed on the target system. This means that data isn’t processed by using an interim transformation in buffered memory.
- Instead, dedicated compute resources must process the data: design computational scalability for data processing

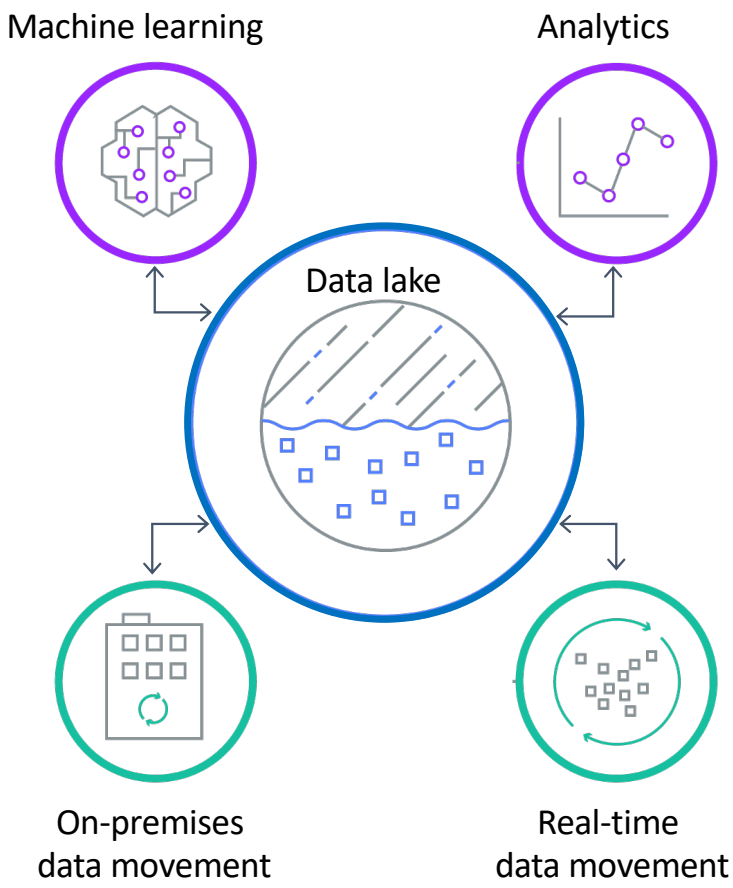
The simplified iterative data pipeline



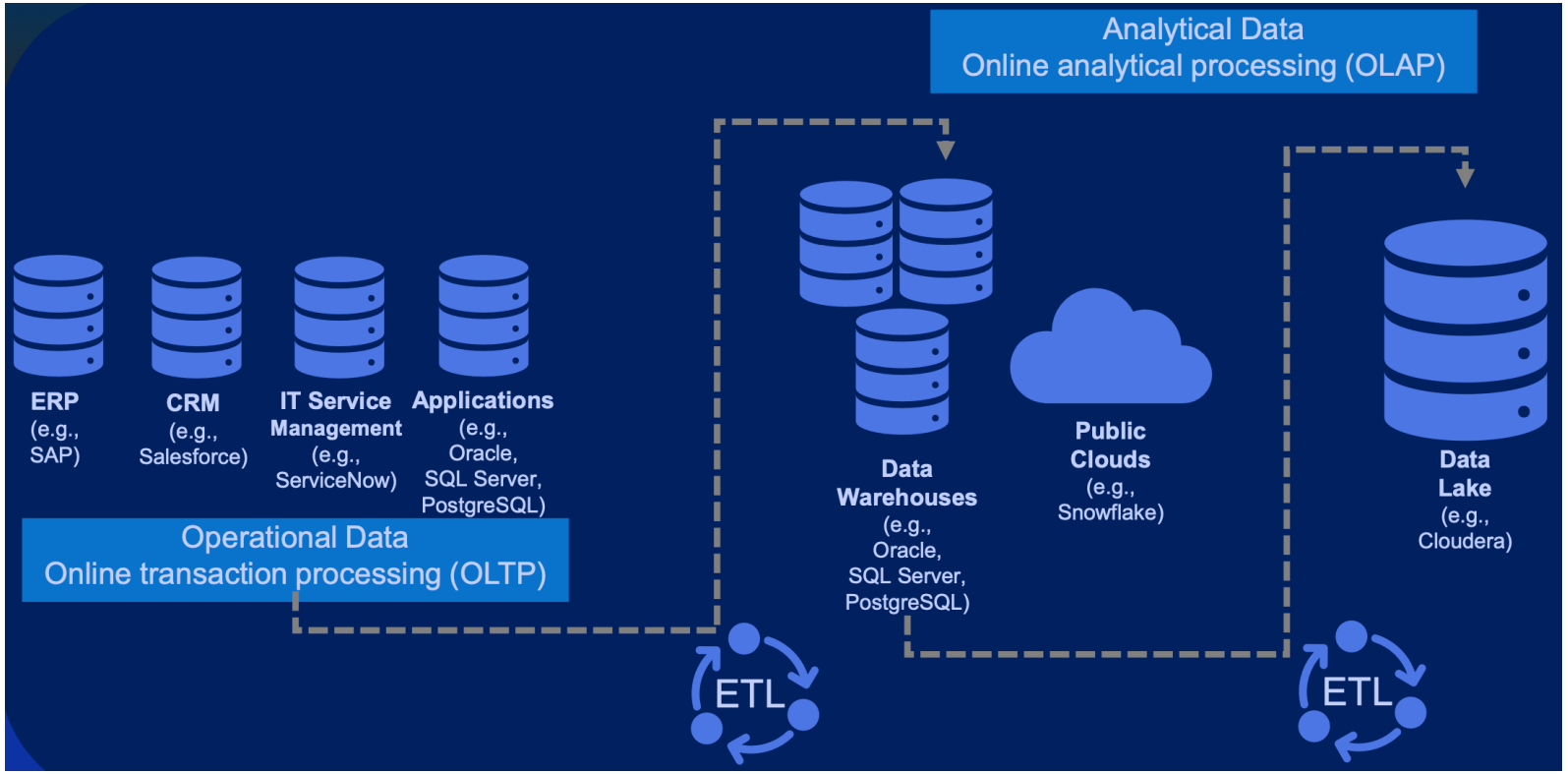
how storage will be used to support your ML pipeline goals?

Data lakes

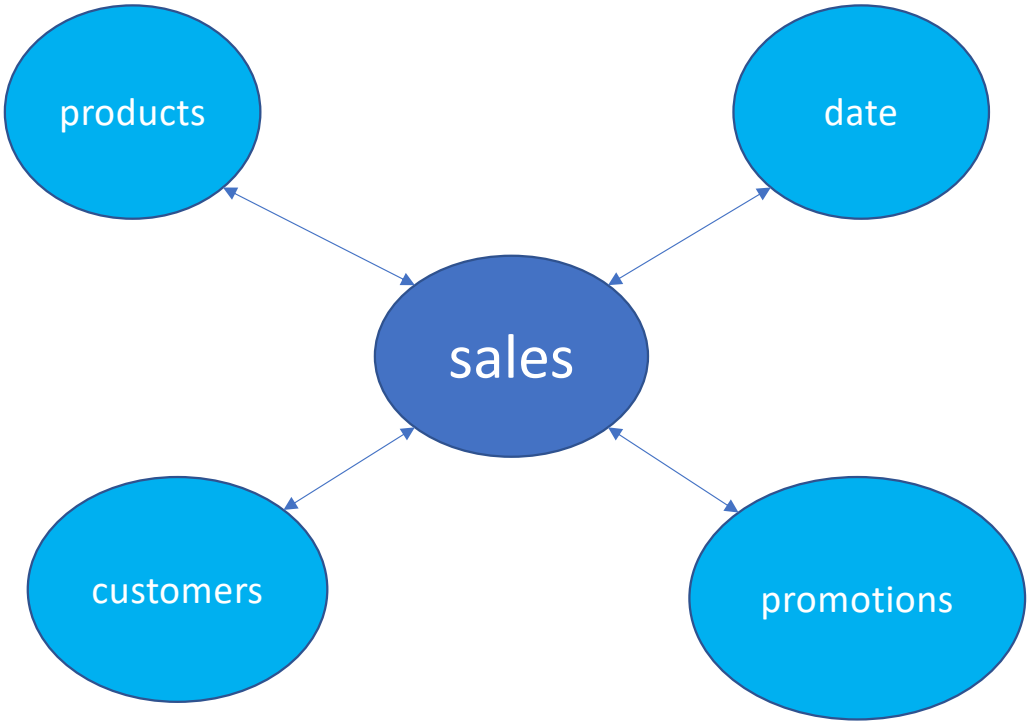
- centralized data repositories
- structured and unstructured data, regardless of scale
- data with original formats
- data without a particular focus on design or usage plans
- the structure of the data or schema is not defined when the data is captured

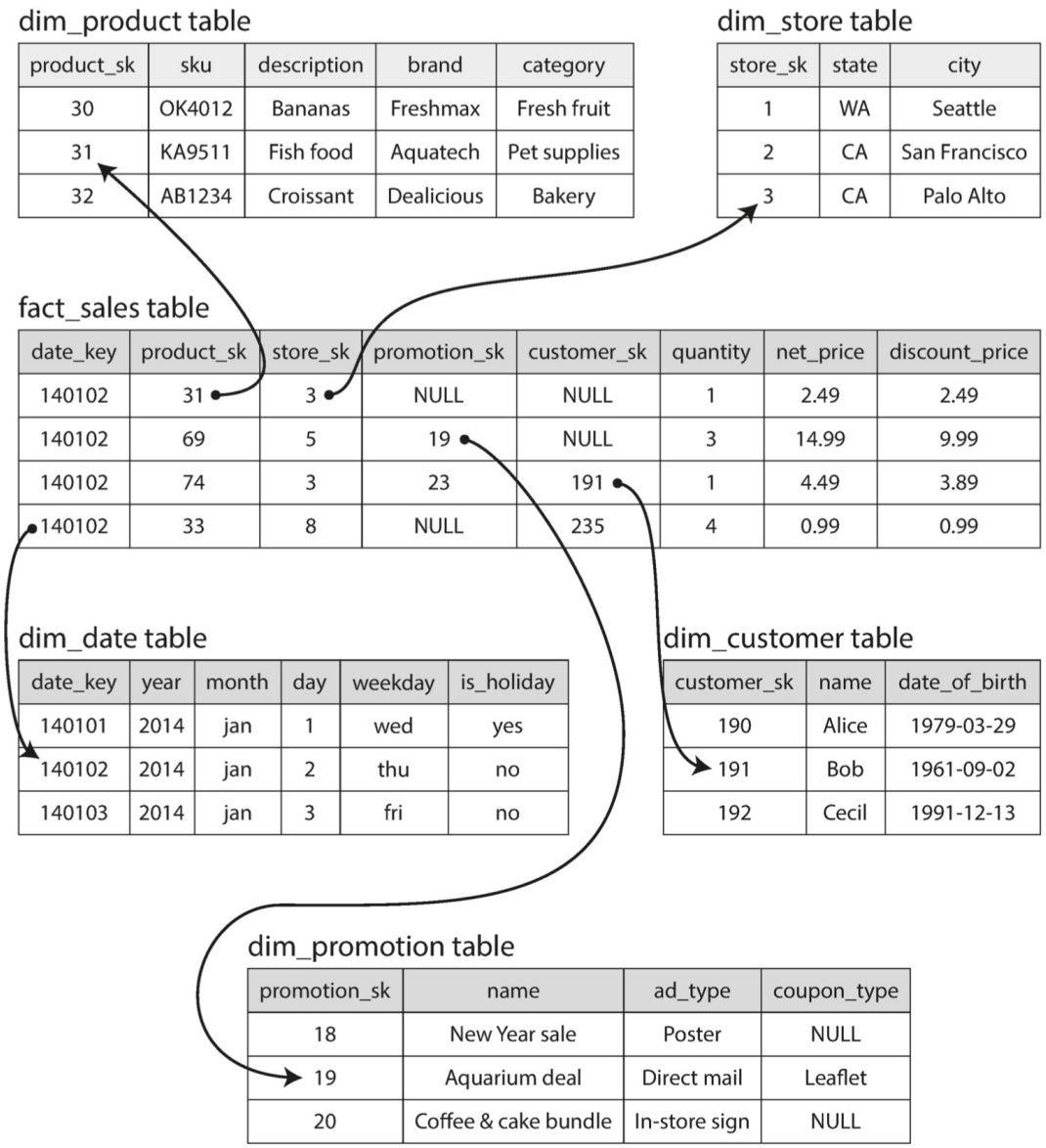


Data warehousing for analytics



Our sales OLAP star-model




dim_product table
product_sk sku description brand category
30 OK4012 Bananas Freshmax Fresh fruit
31 KA9511 Fish food Aquatech Pet supplies
32 AB1234 Croissant Dealicious Bakery
dim_store table
store_sk state city
1 WA Seattle
2 CA San Francisco
3 CA Palo Alto
fact_sales table
date_key product_sk store_sk promotion_sk customer_sk quantity net_price discount_price
140102 31 3 NULL NULL 1 2.49 2.49
140102 69 5 19 NULL 3 14.99 9.99
140102 74 3 23 191 1 4.49 3.89
140102 33 8 NULL 235 4 0.99 0.99
dim_date table
date_key year month day weekday is_holiday
140101 2014 jan 1 wed yes
140102 2014 jan 2 thu no
140103 2014 jan 3 fri no
dim_customer table
customer_sk name date_of_birth
190 Alice 1979-03-29
191 Bob 1961-09-02
192 Cecil 1991-12-13
dim_promotion table
promotion_sk name ad_type coupon_type
18 New Year sale Poster NULL
19 Aquarium deal Direct mail Leaflet
20 Coffee & cake bundle In-store sign NULL

Star schema

- Center of the schema: fact table
- Each element of the fact table represents an event of a particular time

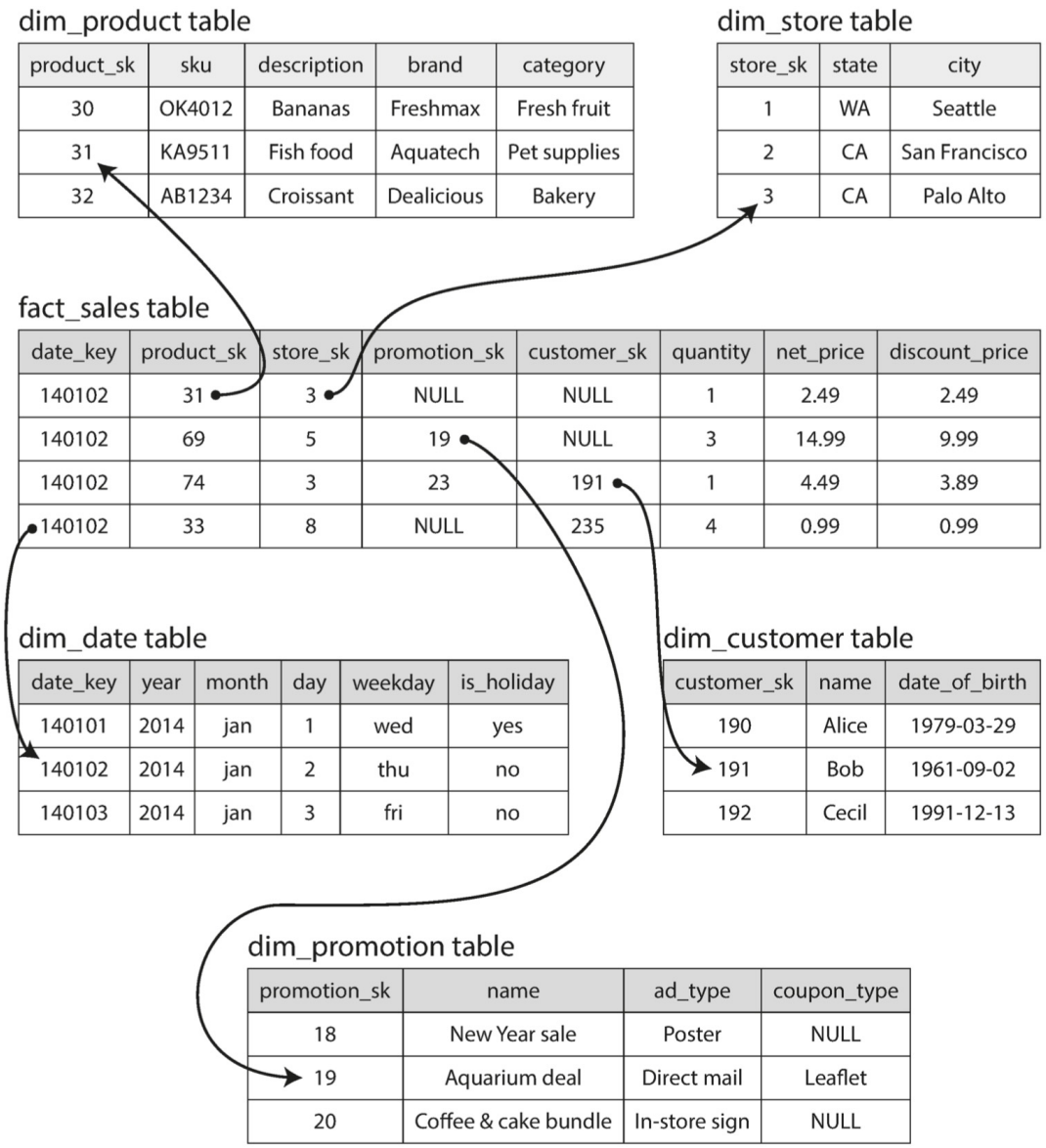
- Example: fact_sales
- Fact element: customer's purchase of a product

Dimension tables to explain data story

- Facts are captured as individual events
- Allows different types of analysis
- Fact table can become very large: petabytes of transaction history
- Each column in fact table makes reference to other tables
- Each dimension represents a way of analysing events history:
  - who
  - what
  - where
  - when
  - how
  - why

Data analyst: build answers to questions

- Each dimension represents a way of analysing events history:
  - **who** bought TV sets during black friday promotion?
  - **what** was most purchased product last night?
  - **where** (which city) is requesting more groceries?
  - **when** are TV sets being purchased (hour)?
  - **how** are promotion and sales related?
  - **why** are product categories in promotion not purchased?


dim_product table
product_sk sku description brand category
30 OK4012 Bananas Freshmax Fresh fruit
31 KA9511 Fish food Aquatech Pet supplies
32 AB1234 Croissant Dealicious Bakery
dim_store table
store_sk state city
1 WA Seattle
2 CA San Francisco
3 CA Palo Alto
fact_sales table
date_key product_sk store_sk promotion_sk customer_sk quantity net_price discount_price
140102 31 3 NULL NULL 1 2.49 2.49
140102 69 5 19 NULL 3 14.99 9.99
140102 74 3 23 191 1 4.49 3.89
140102 33 8 NULL 235 4 0.99 0.99
dim_date table
date_key year month day weekday is_holiday
140101 2014 jan 1 wed yes
140102 2014 jan 2 thu no
140103 2014 jan 3 fri no
dim_customer table
customer_sk name date_of_birth
190 Alice 1979-03-29
191 Bob 1961-09-02
192 Cecil 1991-12-13
dim_promotion table
promotion_sk name ad_type coupon_type
18 New Year sale Poster NULL
19 Aquarium deal Direct mail Leaflet
20 Coffee & cake bundle In-store sign NULL

Data warehouse: three service tiers

1. At the top is a front-end client, which presents your results through reporting, analysis, and data mining tools.
2. The middle tier is made up of an analytics engine, which is used to access and analyze the data.
3. The third tier consists of the database server, where your data is loaded and stored.
   - frequently accessed data: fast storage like SSD drives
   - infrequently accessed: low-cost object storage, such as Amazon S3
   - data can be automatically shifted between the two storage types based on how frequently it is accessed.

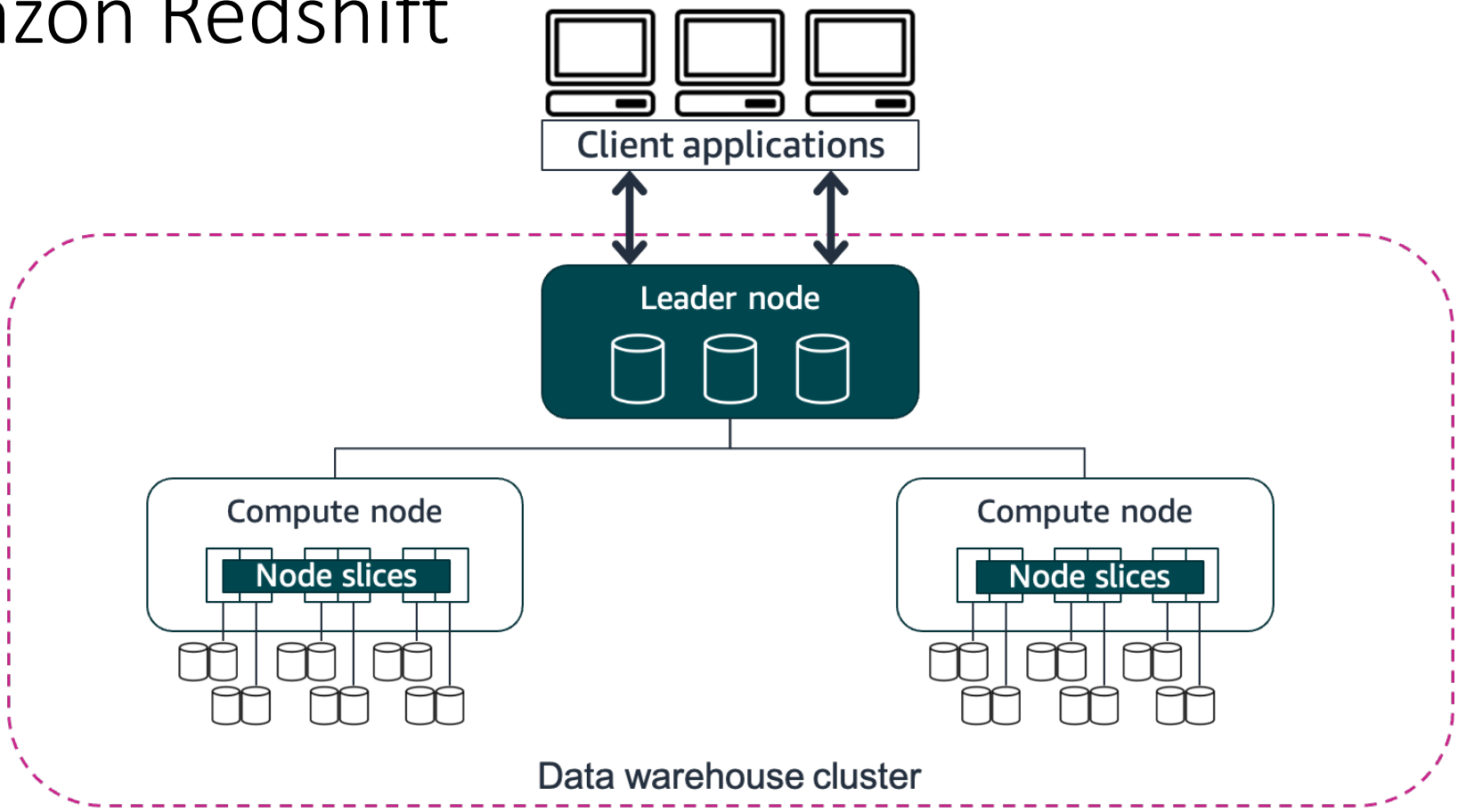
Amazon Redshift

- Provides a cloud-based data warehouse solution
- Is a fully managed service
- Supports near real-time data analysis
- Uses columnar storage
- Collections of computing resources: nodes

Redshift clusters, nodes and databases

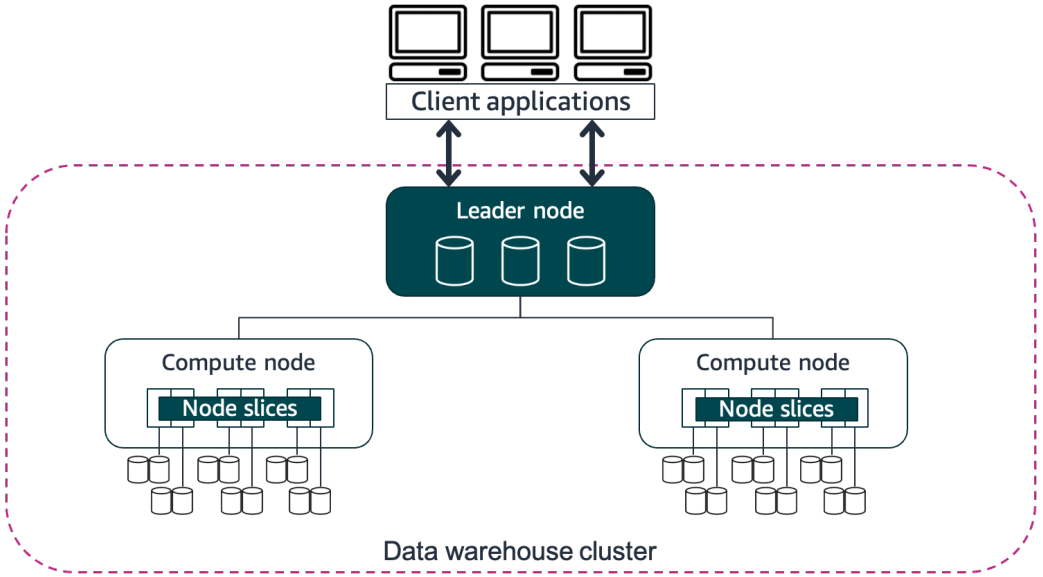
- The service consists of collections of computing resources called *nodes*.
- Nodes are organized into *clusters* that run an Amazon Redshift engine and contain one or more column-oriented databases.
- Each cluster has a **leader node** and **one or more compute nodes**.
- The leader node receives queries from client applications, and then parses the queries and develops query execution plans.
- The leader node then coordinates the parallel processing of the query plans with the compute nodes and aggregates the intermediate results.
- Finally, the results are returned to the client applications.

Example architecture: Data warehouse in Amazon Redshift



Data warehouse cluster: leader and two compute nodes

- compute node: returns the intermediate results of the query back to the leader node for aggregation.
- has dedicated CPU, memory, and attached disk storage based on node type
- is partitioned into *node slices*.
  - have a portion of the compute node's memory and disk space.
  - work in parallel to complete operations that the leader node assigned to the computer node.



Why separating data warehouse from OLTP?

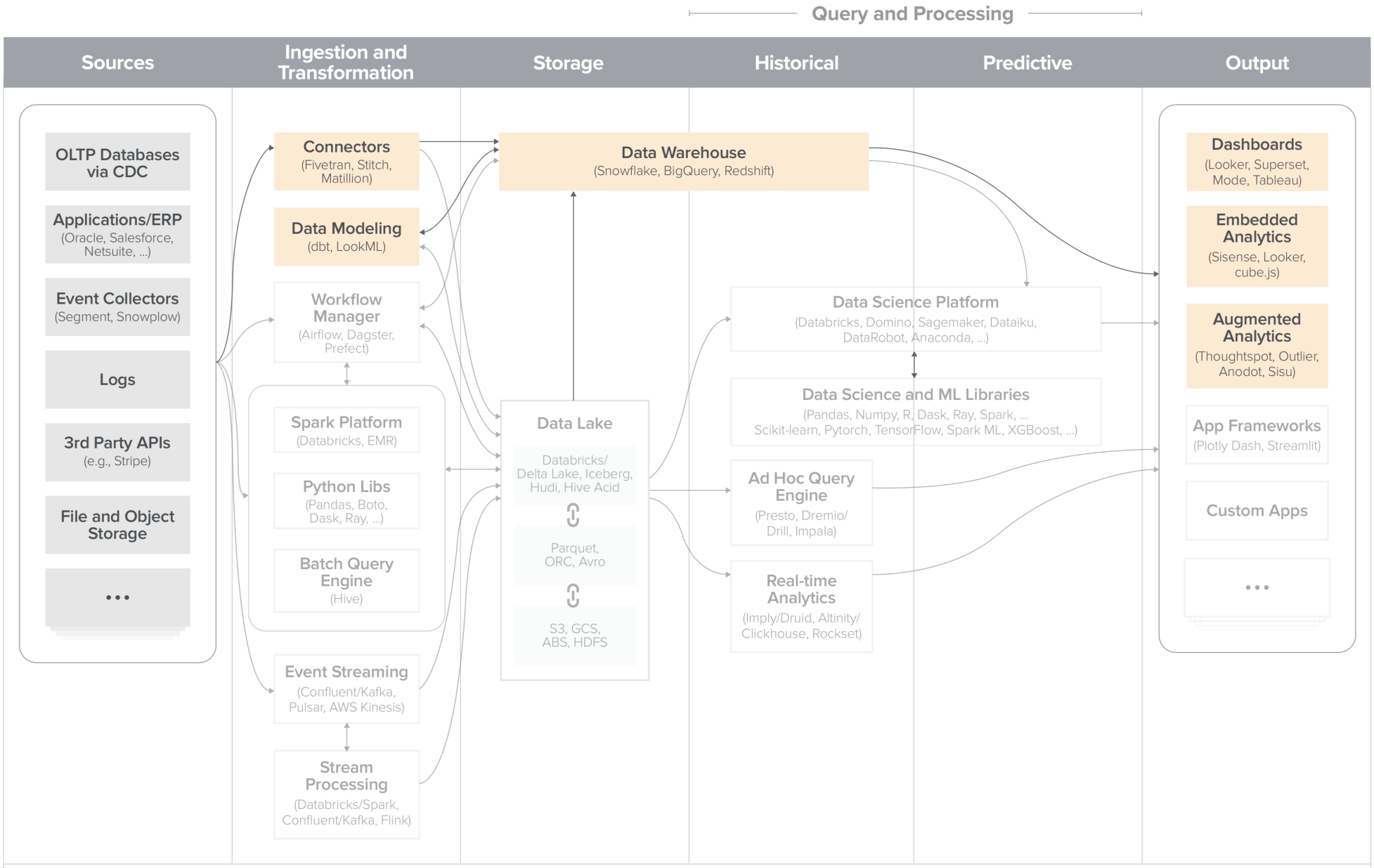
- Optimize for analytic access patterns
- Indexing algorithms for OLTP not good for analytic queries

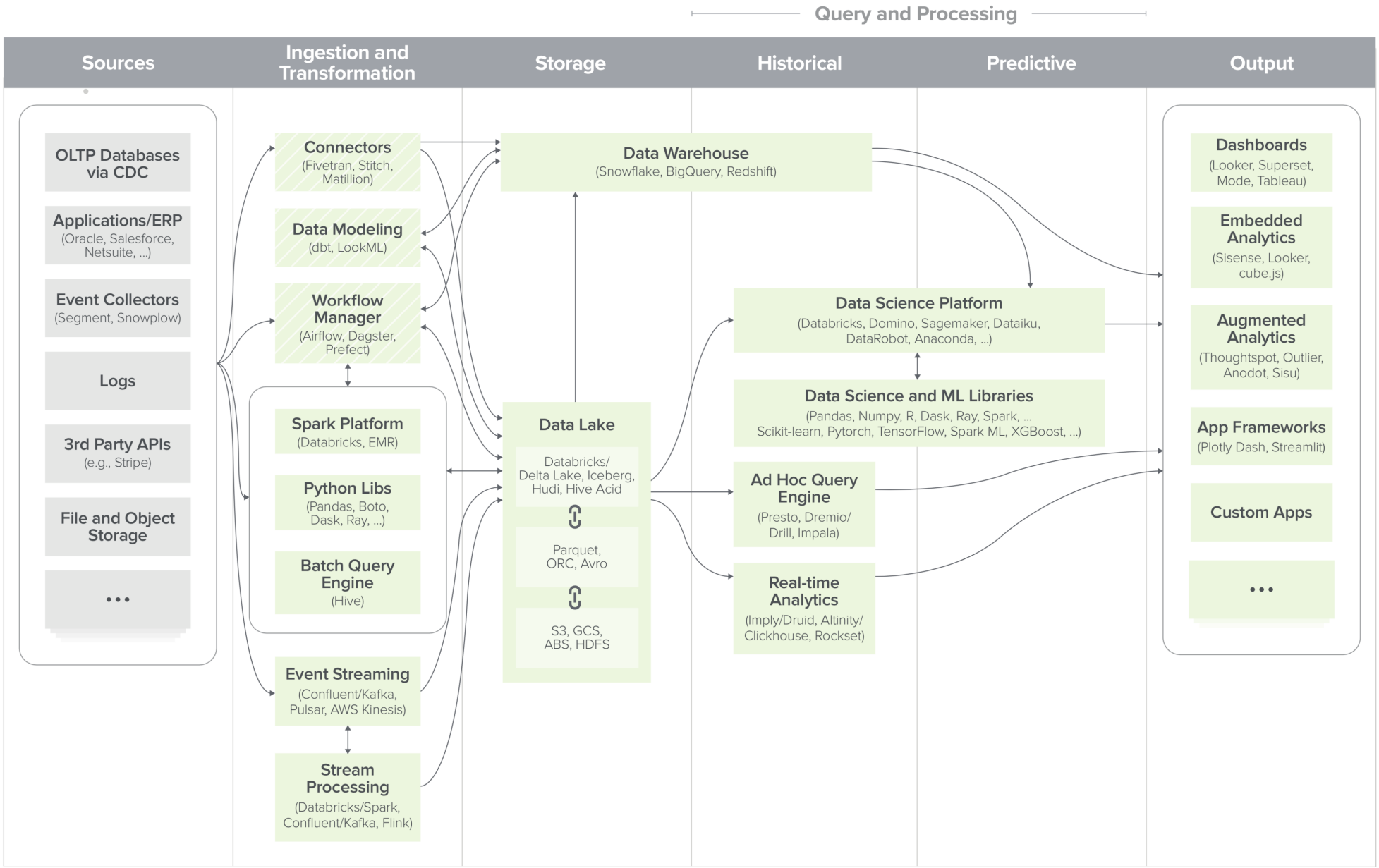
- Data warehouse vendors: Teradata, Vertica, SAP HANA, ParAccel, AWS Redshift
- Open source Hadoop projects: Apache Hive, Spark, Cloudera Impala, Facebook Presto, Apache Drill

Emergent architectures for data infrastructure

- Modern Business Intelligence
- Complex systems built around data
- Value of the system is the result of the analysis
- Infrastructure
- Tools

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

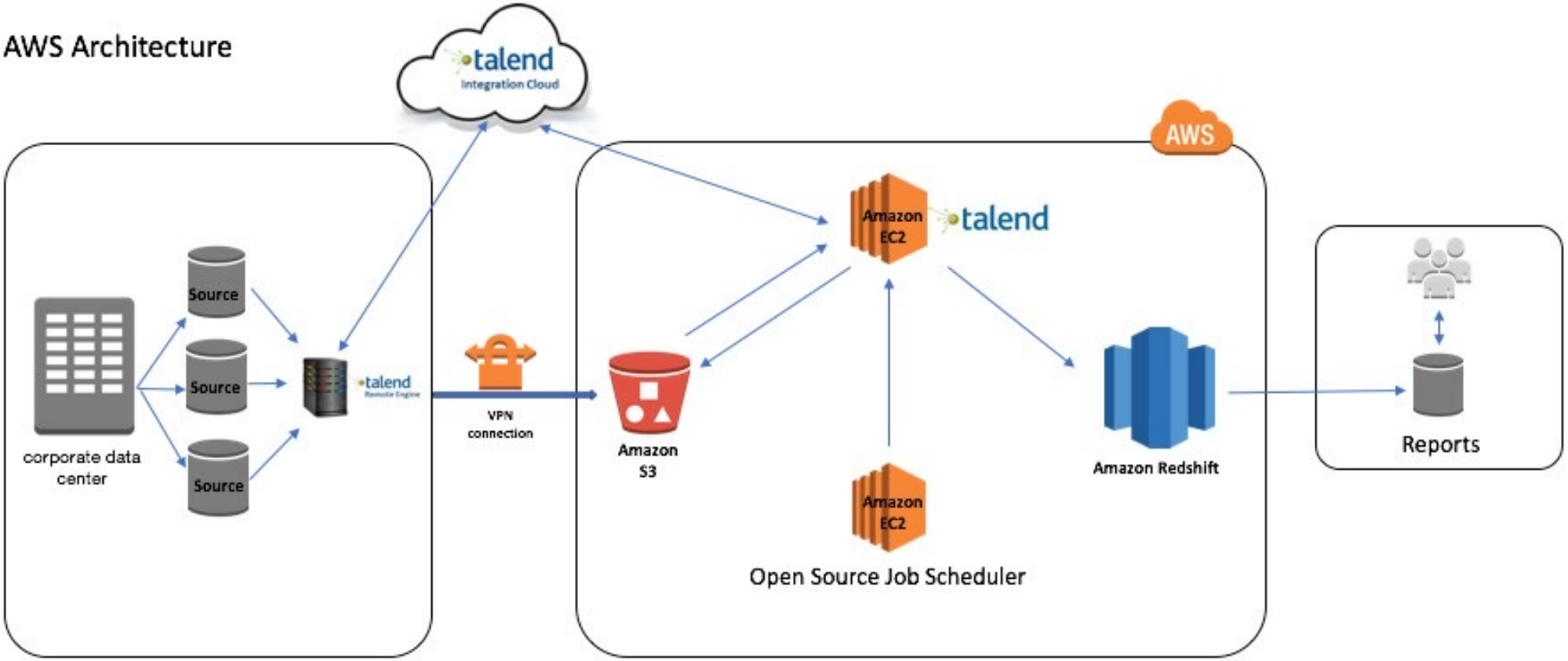

Query and Processing
Sources
Ingestion and Transformation
Storage
Historical
Predictive
Output
OLTP Databases via CDC
Applications/ERP
(Oracle, Salesforce, Netsuite, ...)
Event Collectors
(Segment, Snowplow)
Logs
3rd Party APIs
(e.g., Stripe)
File and Object Storage
...
Connectors
(Fivetran, Stitch, 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)
Data Warehouse
(Snowflake, BigQuery, Redshift)
Data Lake
Databricks/ Delta Lake, Iceberg, Hudi, Hive Acid
Parquet, ORC, Avro
S3, GCS, ABS, HDFS
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, ...)
Ad Hoc Query Engine
(Presto, Dremio/ Drill, Impala)
Real-time Analytics
(Imply/Druid, Altinity/ Clickhouse, Rockset)
Dashboards
(Looker, Superset, Mode, Tableau)
Embedded Analytics
(Sisense, Looker, cube.js)
Augmented Analytics
(Thoughtspot, Outlier, Anodot, Sisu)
App Frameworks
(Plotly Dash, Streamlit)
Custom Apps
...


Query and Processing
Sources
Ingestion and Transformation
Storage
Historical
Predictive
Output
OLTP Databases via CDC
Applications/ERP
(Oracle, Salesforce, Netsuite, ...)
Event Collectors
(Segment, Snowplow)
Logs
3rd Party APIs
(e.g., Stripe)
File and Object Storage
...
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)
Data Warehouse
(Snowflake, BigQuery, Redshift)
Data Lake
Databricks/ Delta Lake, Iceberg, Hudi, Hive Acid
Parquet, ORC, Avro
S3, GCS, ABS, HDFS
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, ...)
Ad Hoc Query Engine
(Presto, Dremio/ Drill, Impala)
Real-time Analytics
(Imply/Druid, Altinity/ Clickhouse, Rockset)
Dashboards
(Looker, Superset, Mode, Tableau)
Embedded Analytics
(Sisense, Looker, cube.js)
Augmented Analytics
(Thoughtspot, Outlier, Anodot, Sisu)
App Frameworks
(Plotly Dash, Streamlit)
Custom Apps
...

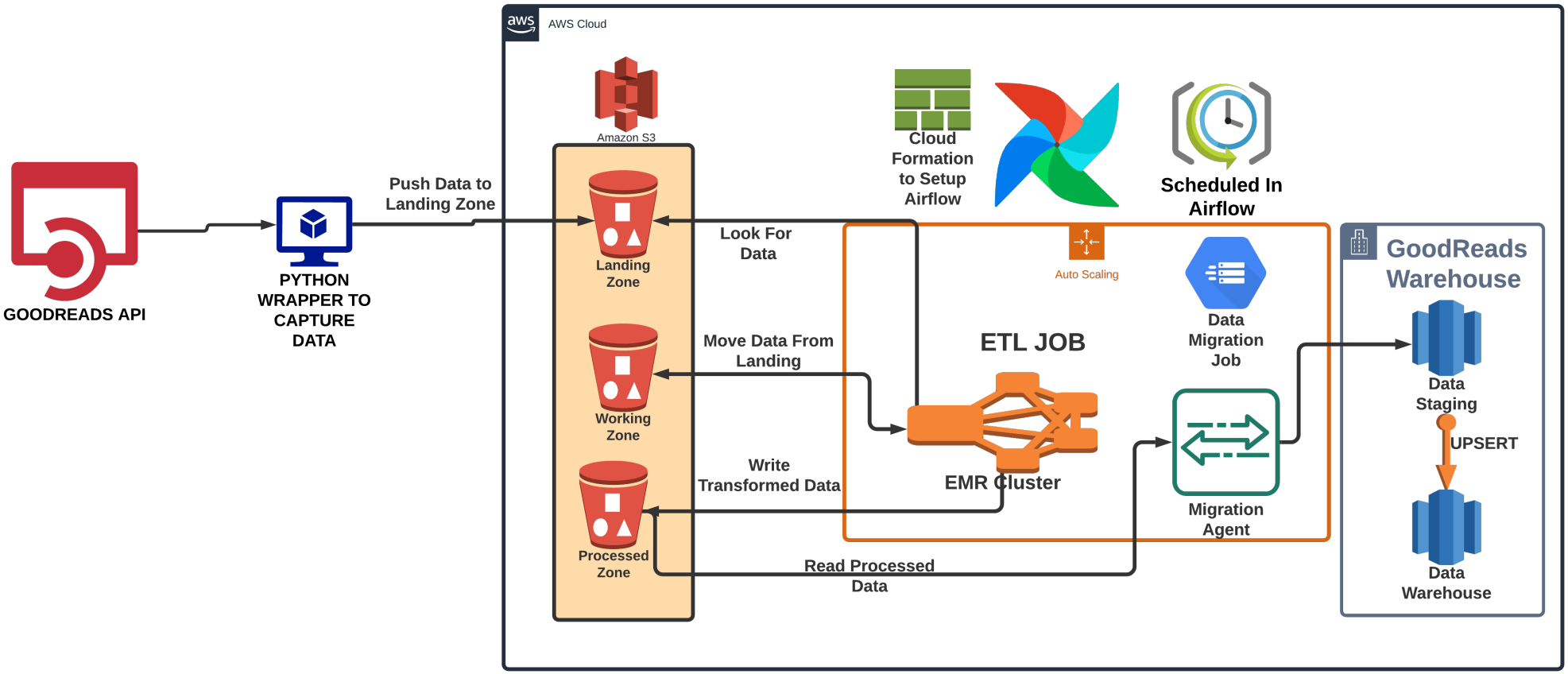
On Prem → Cloud Data Warehouse	Data warehouses are moving to the cloud with increased flexibility, scale, and ease of use—allowing any company to be a data company	snowflake, Google Big Query
Hadoop → Next-gen Data Lakes	Data lakes and related systems are becoming more performant and reliable, adding RDBMS-like features including ACID transactions and interactive SQL queries	databricks, presto
ETL → ELT	Brittle ETL processes (extract-transform-load) are being replaced with more flexible and consistent ELT pipelines (extract-load-transform)	Fivetran, dbt
Workflow Manager → Dataflow Automation	Data flow automation systems are helping to orchestrate thousands of data pipelines with a cleaner abstraction and modern executor integrations	PREFECT, DAGSTER, Apache Airflow
Analyst Teams → Self-serve Insights	Reporting, dashboarding, and automated analysis tools are becoming more available to non-technical users	Looker, Superset
Endpoint Protection → Global Data Governance	Data security and privacy measures (e.g., access controls) are becoming centralized on the data platform as use of data is increasingly regulated and user endpoints are harder to protect	Collibra, PRIVACERA

Cloud Data Warehouse

Microsoft Azure case



https://aws.amazon.com/blogs/database/using-amazon-redshift-for-fast-analytical-reports/



https://github.com/san089/goodreads_etl_pipeline

The Data Journey

Data Ingestion

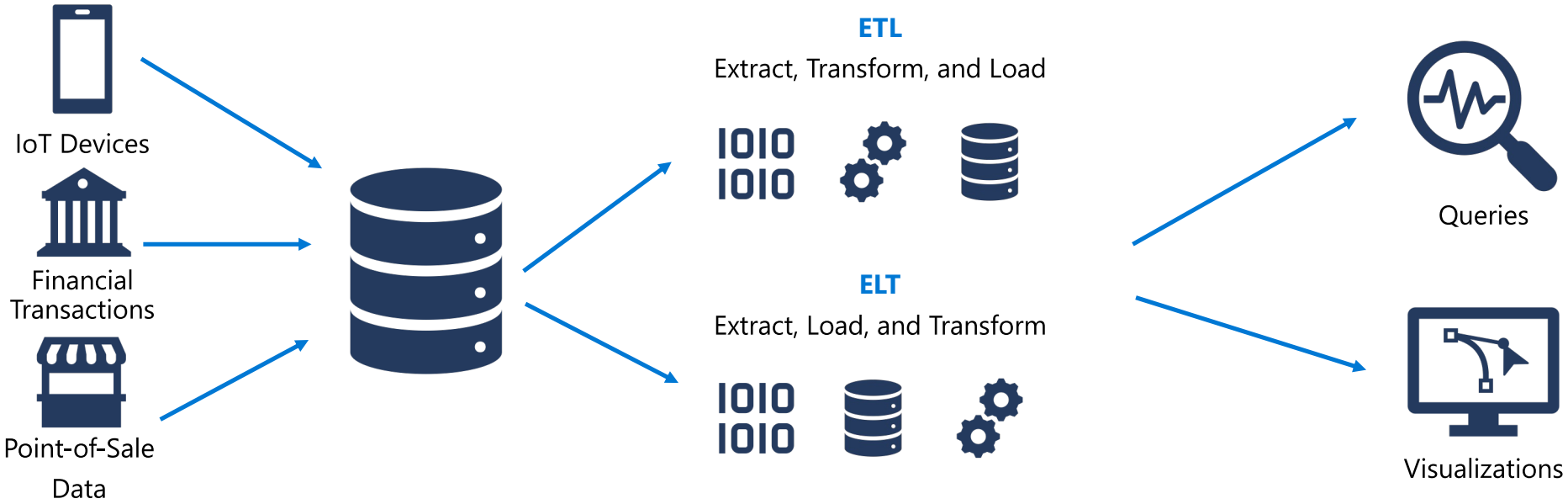
The process of obtaining and importing data for immediate use or storage in a database

Data Processing

Takes the data in its raw form, cleans it, and converts it into a more meaningful format

Data Visualization

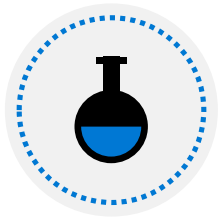
Query the data and create graphical representations of information and data



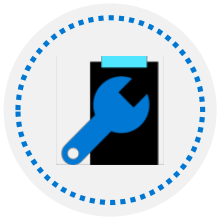
Explore data analytics

Descriptive

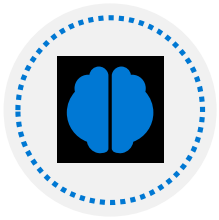
Diagnostic



Predictive



Prescriptive



Cognitive

Azure services for data warehouse

Azure Data Factory

Azure Data Lake

Azure Databricks

Spark-based

Azure HDInsight

Spark, Hive, Kafka, Storm

What is Azure Data Factory?

Azure Data Factory is described as a data integration service.

- Retrieves data from more than one data source and converts it.
- Filters out noise to keep interesting data
- Work is defined as a pipeline operation – runs continuously as data is received

What is Azure Data Lake Storage?

Azure Data Lake Storage is a repository of data for your modern data warehouse.

- Organizes data into directories for improved file access.
- Supports POSIX and RBAC permissions.
- Compatible with the Hadoop Distributed File System

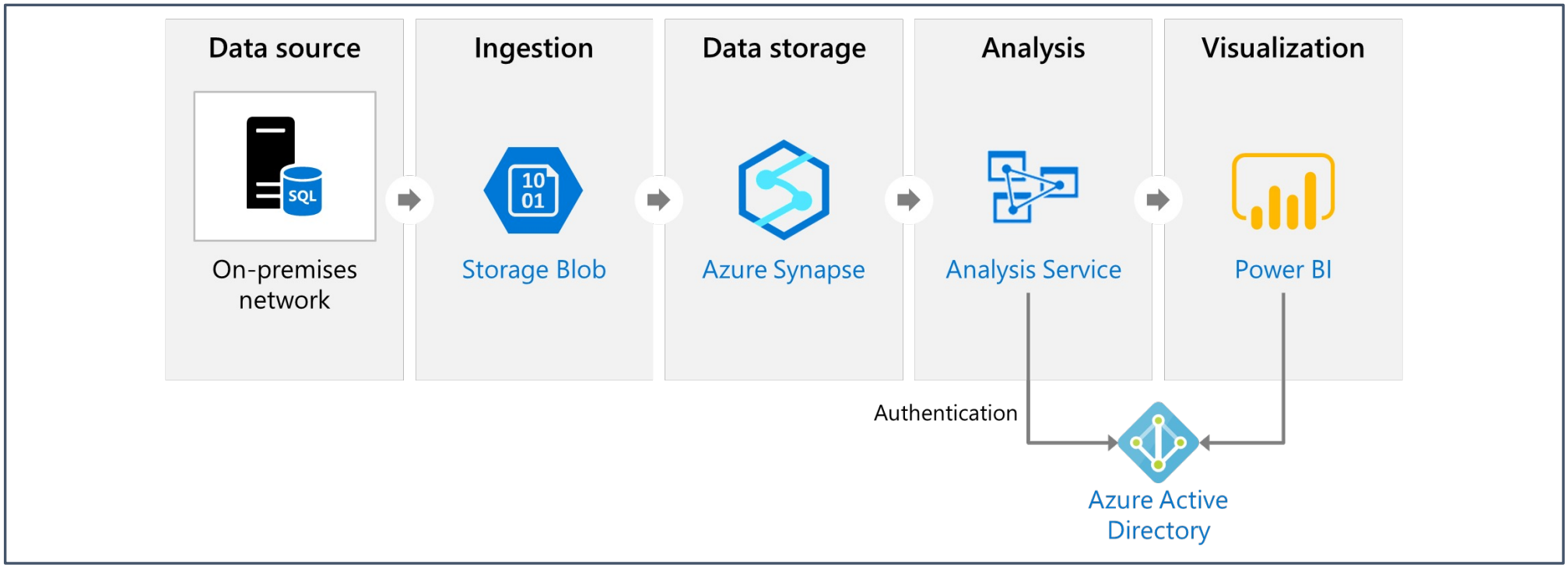
What is Azure Databricks?

Azure Databricks is an **Apache Spark-based platform** that provides big data processing and streaming.

- Simplifies the provisioning and collaboration of Apache spark-based analytical solutions.
- Utilizes the security capabilities of Azure.
- Integrates with a variety of Azure data platform services and Power BI.

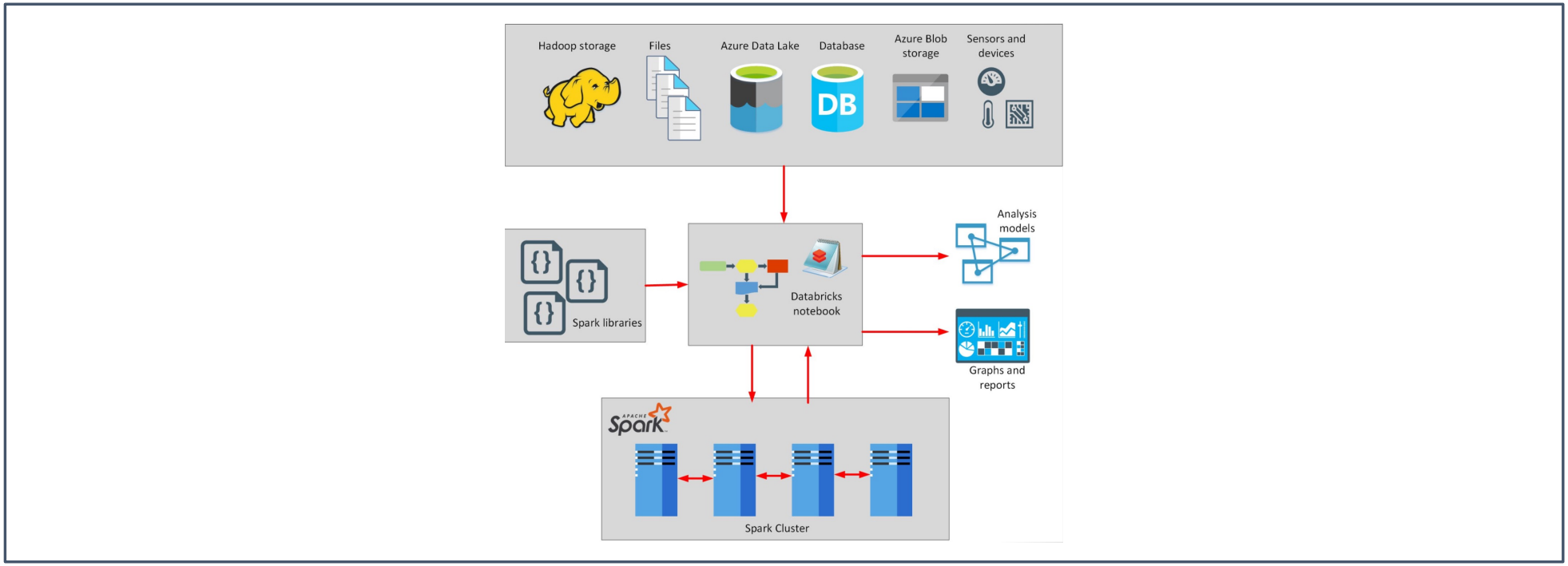
What is Azure Analysis Services?

Azure Analysis Services builds tabular models to support online analytical processing (OLAP) queries.



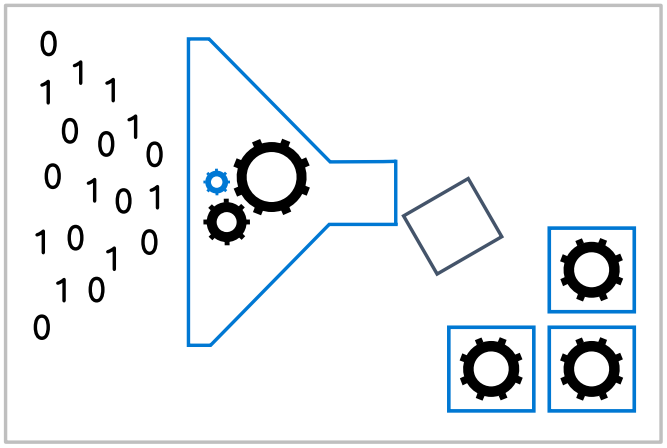
What is Azure HDInsight?

Azure HDInsight is a big data processing services which allows you to use open-source libraries on the one platform, in an Azure environment.



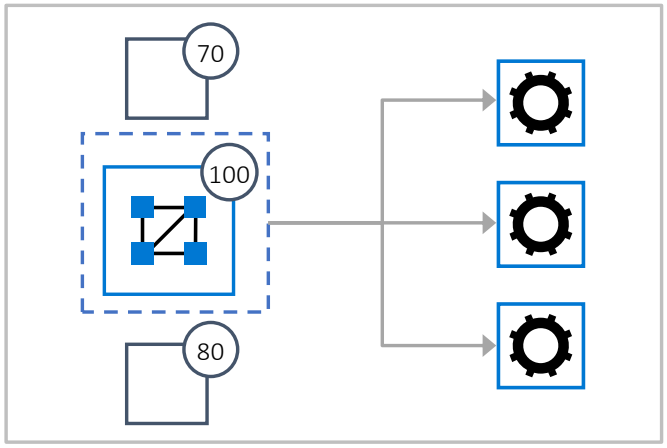
Data ingestion processing

Data Factory



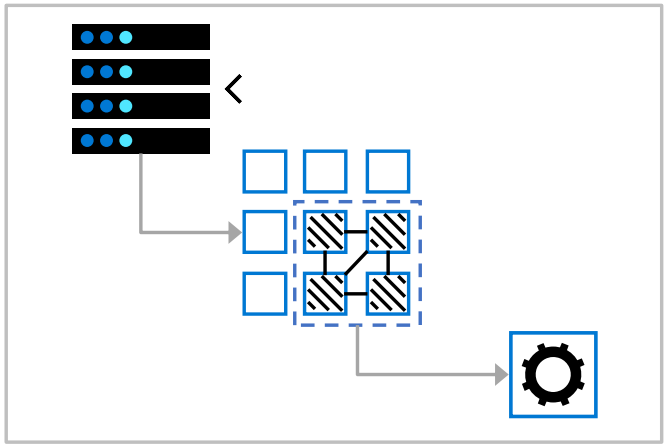
Heterogenous sources

PolyBase



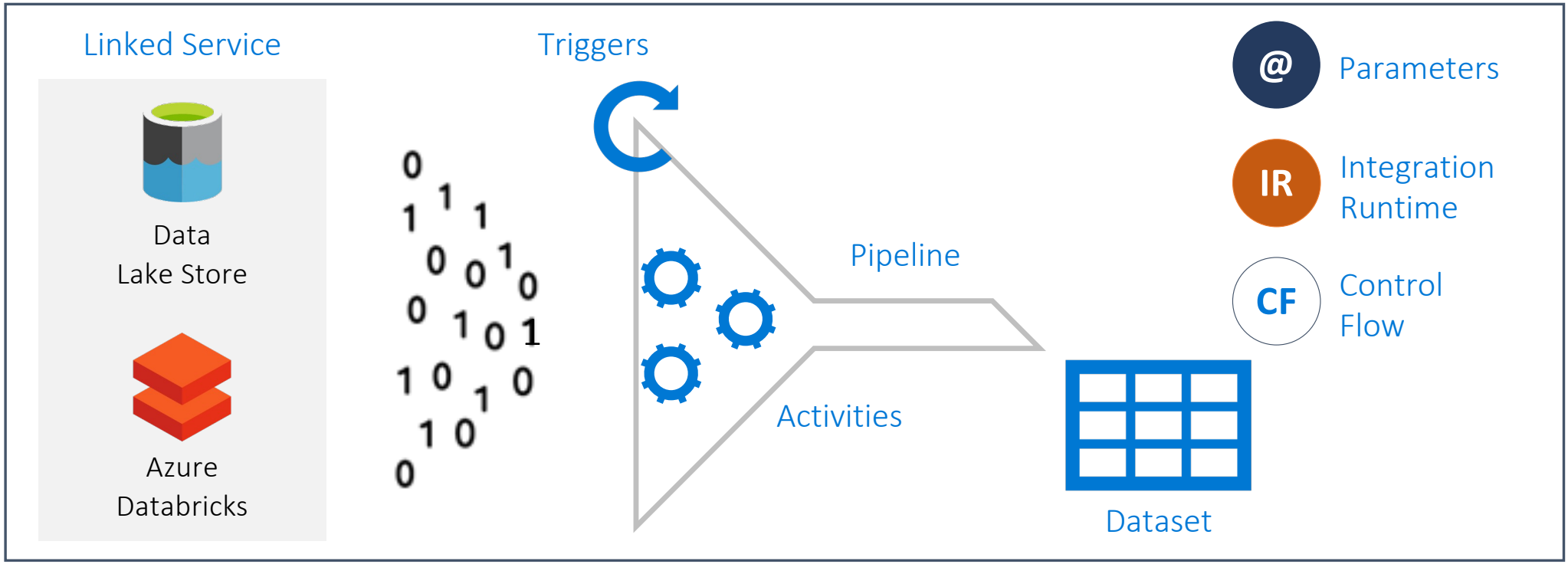
Extract from SQL Server

SSIS

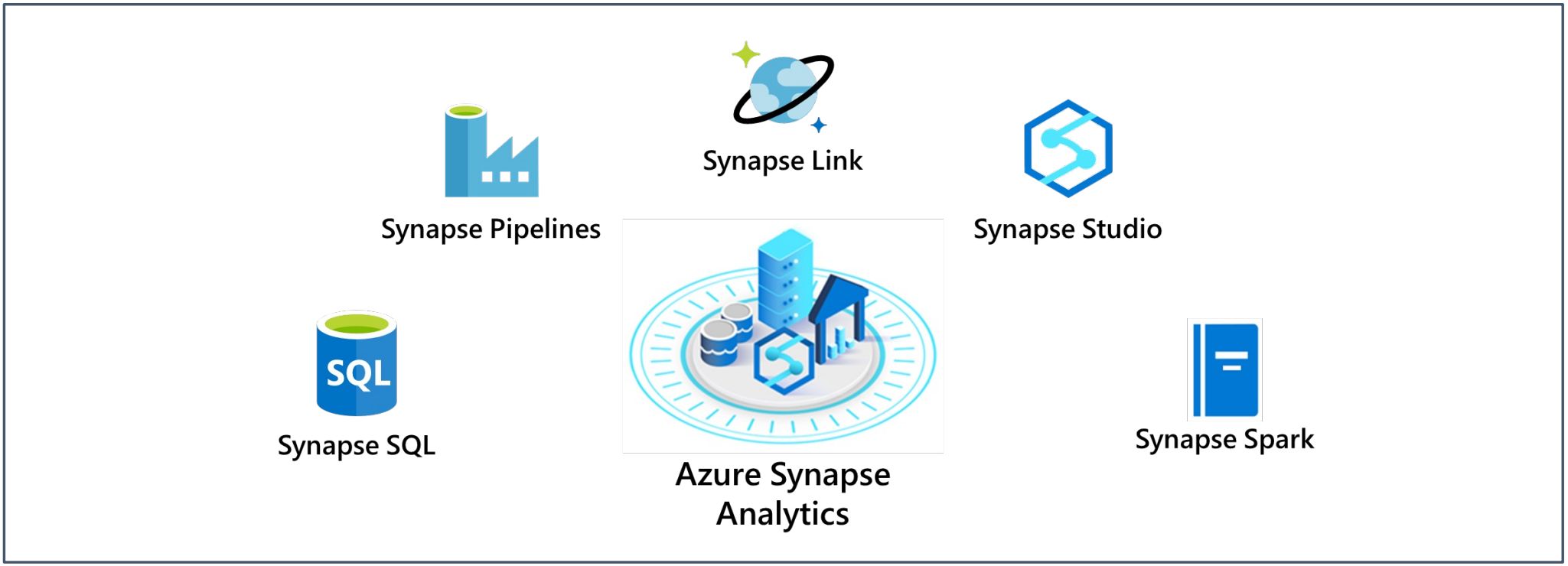


Heterogenous

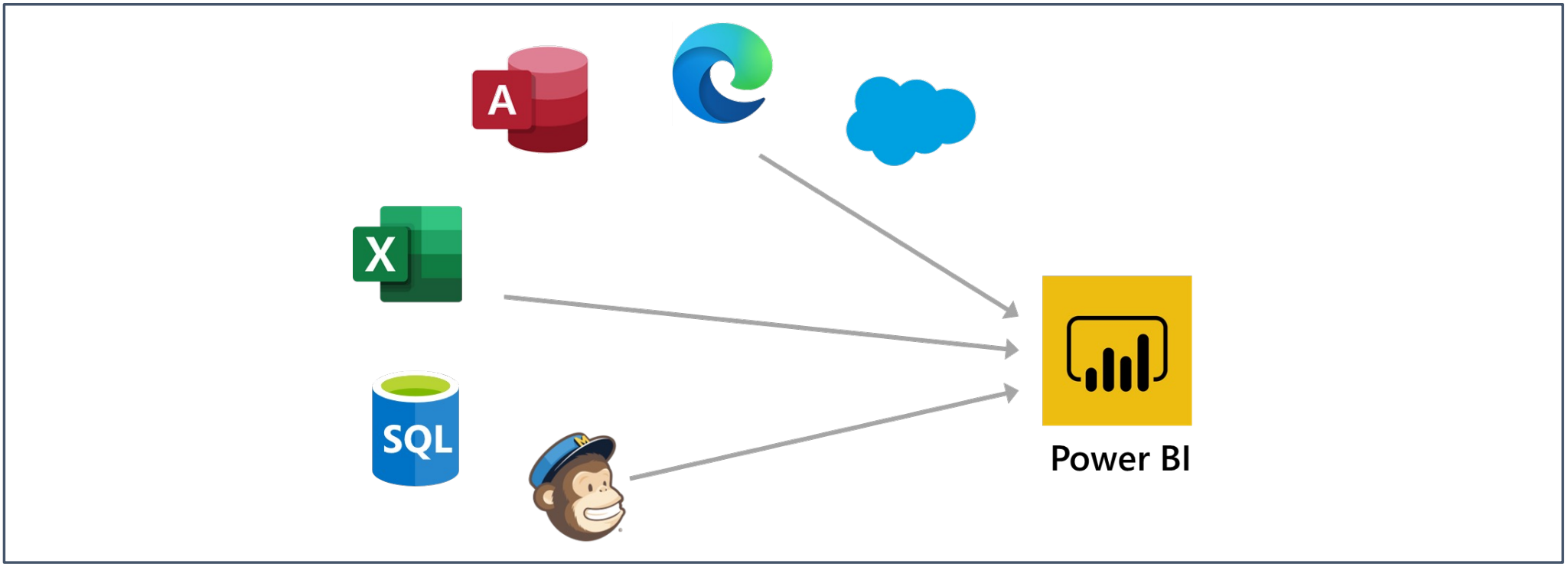
Components of Azure Data Factory



Explore Azure Synapse Analytics



What is Power BI?

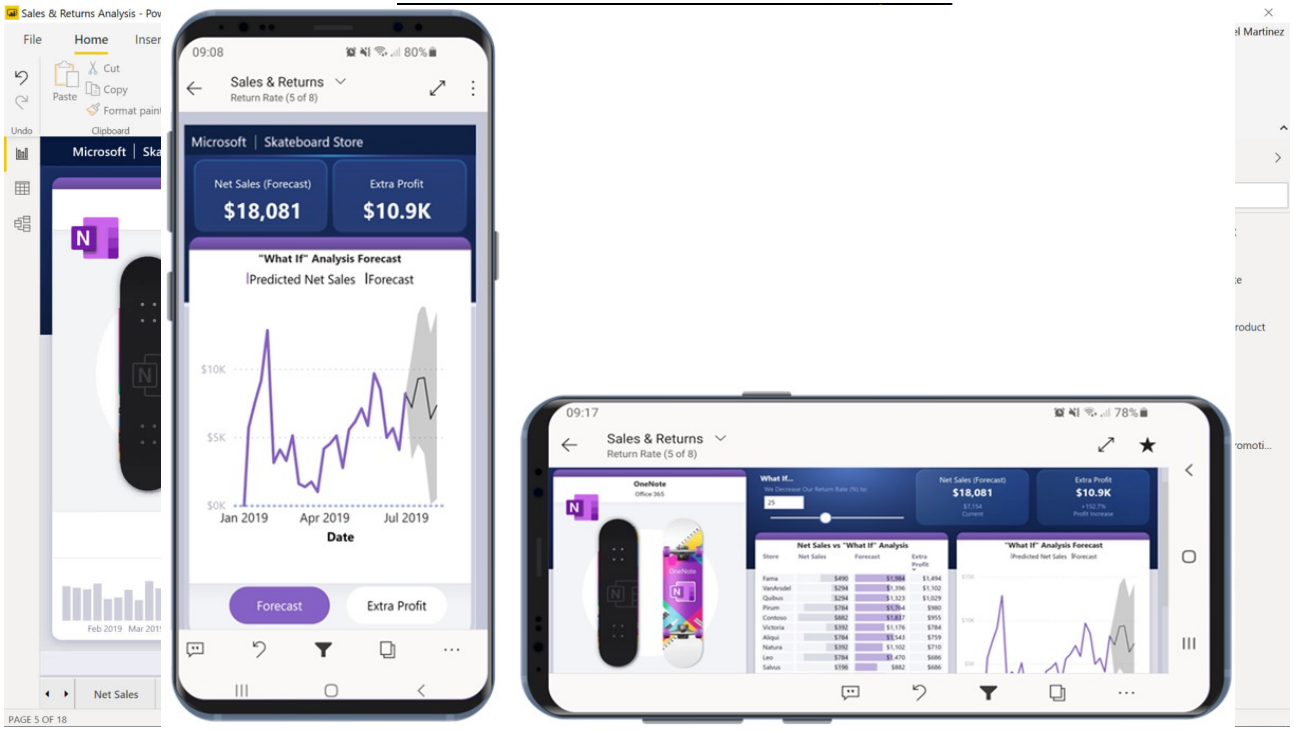


How can you use Power BI?

Power BI Desktop

Power BI Service

Power BI Mobile



The anatomy of a Power BI app

Visualizations

Datasets

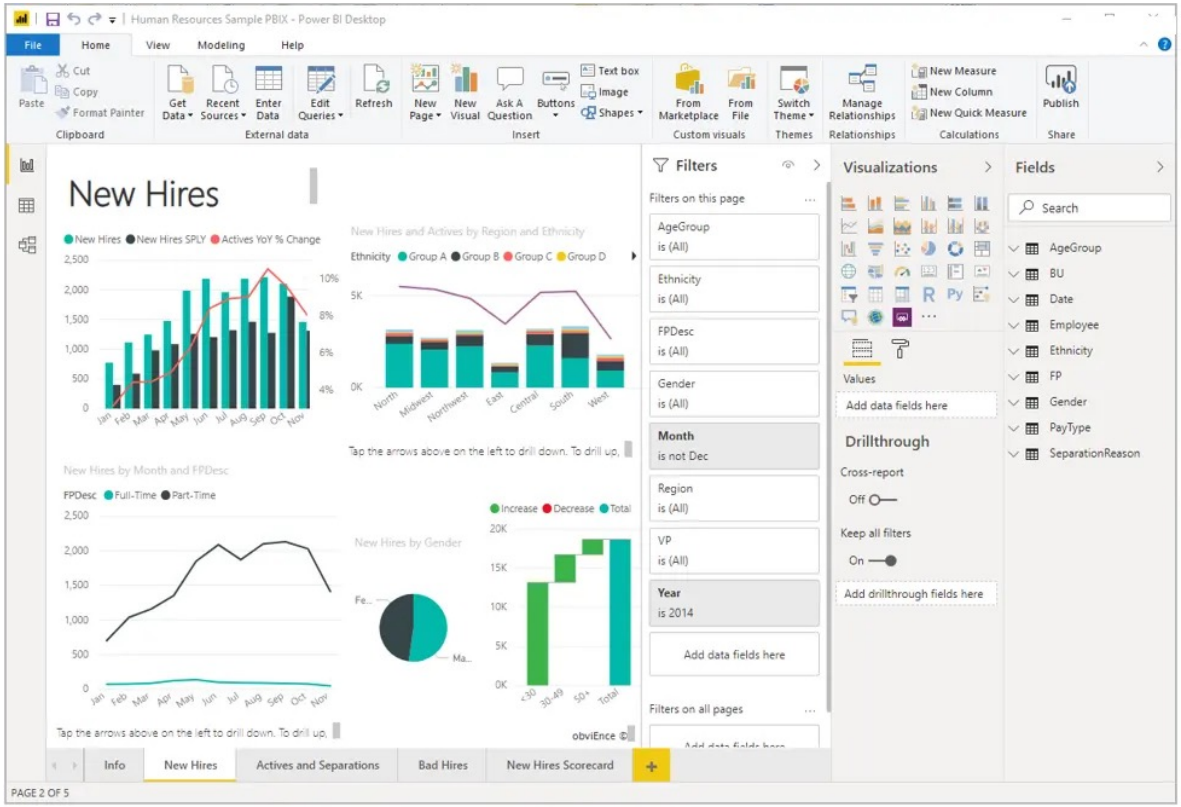
C2132 | 2

	B	C	D	E	F	G	H
1	Year	Month	Month Name	Calendar Month	Births	Births Per Day	Births (Normalized)
2119	2004	1	January	1/1/2004	2,937	94.7	2842
2120	2004	2	February	2/1/2004	2,824	97.4	2921
2121	2004	3	March	3/1/2004	3,128	100.9	3027
2122	2004	4	April	4/1/2004	2,896	96.5	2896
2123	2004	5	May	5/1/2004	3,008	97.0	2911
2124	2004	6	June	6/1/2004	3,047	101.6	3047
2125	2004	7	July	7/1/2004	2,981	96.2	2885
2126	2004	8	August	8/1/2004	3,079	99.3	2980
2127	2004	9	September	9/1/2004	3,219	107.3	3219
2128	2004	10	October	10/1/2004	3,547	114.4	3433
2129	2004	11	November	11/1/2004	3,365	112.2	3365
2130	2004	12	December	12/1/2004	3,143	101.4	3042
2131	2005	1	January	1/1/2005	2,921	94.2	2827
2132	2005	2	February	2/1/2005	2,699	96.4	2892
2133	2005	3	March	3/1/2005	3,024	97.5	2926
2134	2005	4	April	4/1/2005	3,037	101.2	3037
2135	2005	5	May	5/1/2005	3,231	104.2	3127
2136	2005	6	June	6/1/2005	3,163	105.4	3163
2137	2005	7	July	7/1/2005	3,119	100.6	3018
2138	2005	8	August	8/1/2005	3,156	101.8	3054
2139	2005	9	September	9/1/2005	3,439	114.6	3439

The anatomy of a Power BI app

Visualizations

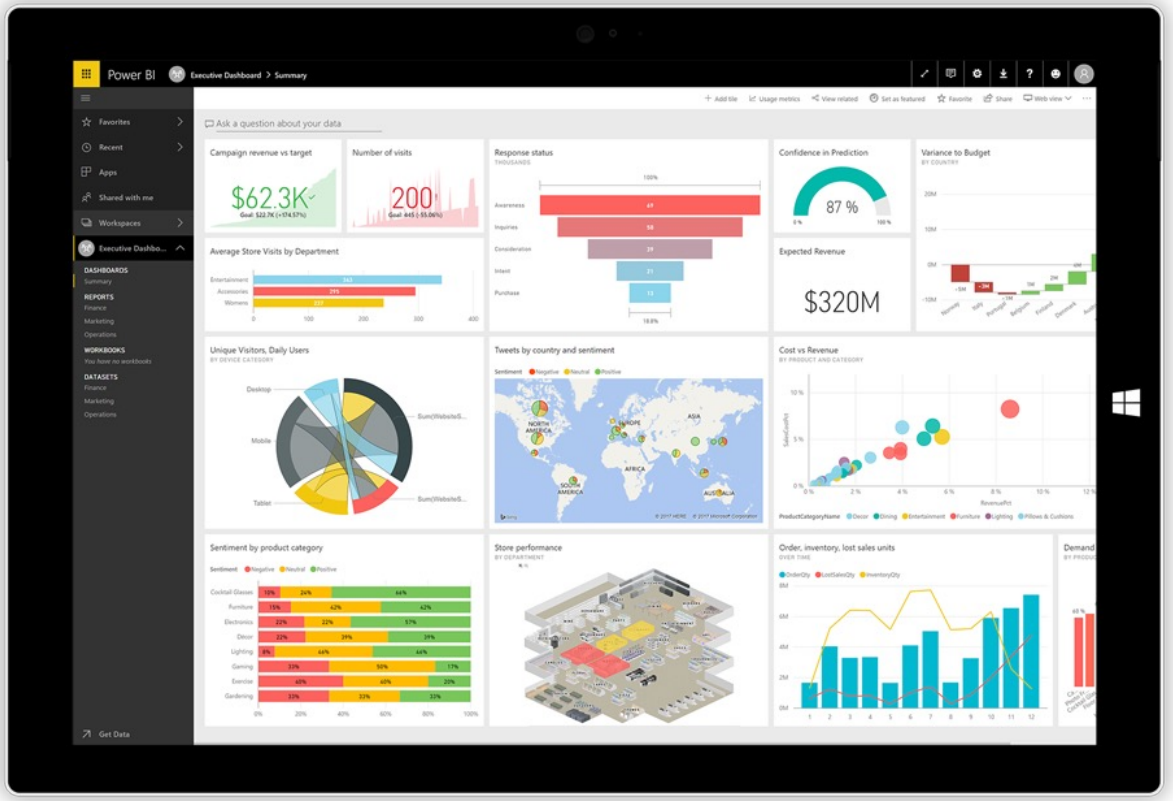
Reports



The anatomy of a Power BI app

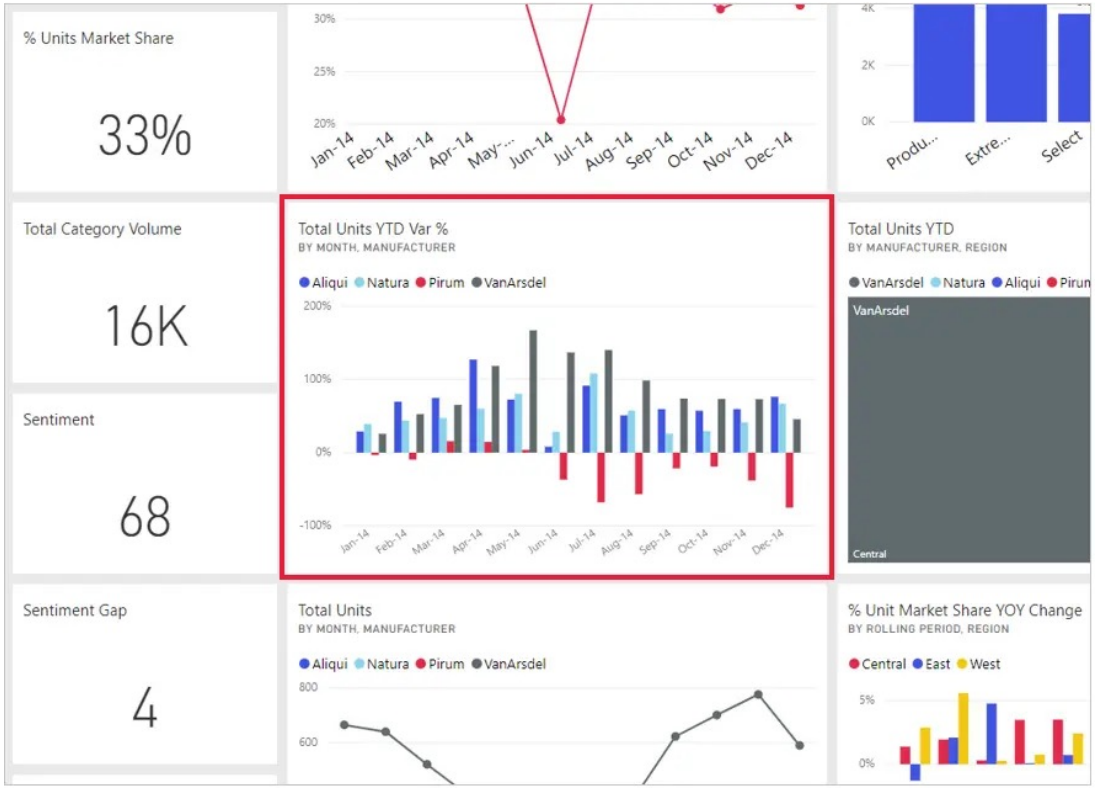
Visualizations

Dashboards



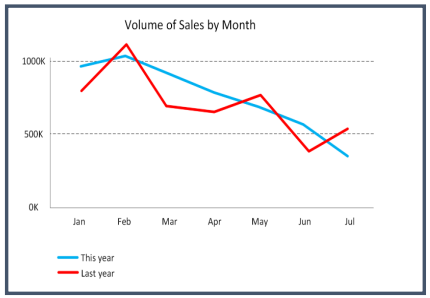
The anatomy of a Power BI app

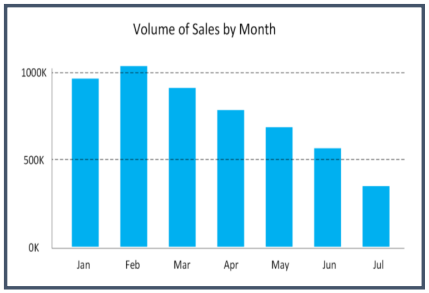
Dashboards Tile

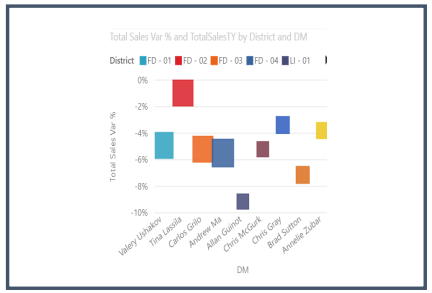


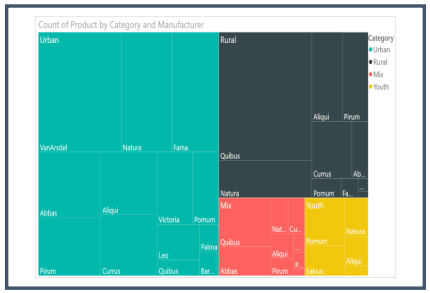
The anatomy of a Power BI app

Tiles





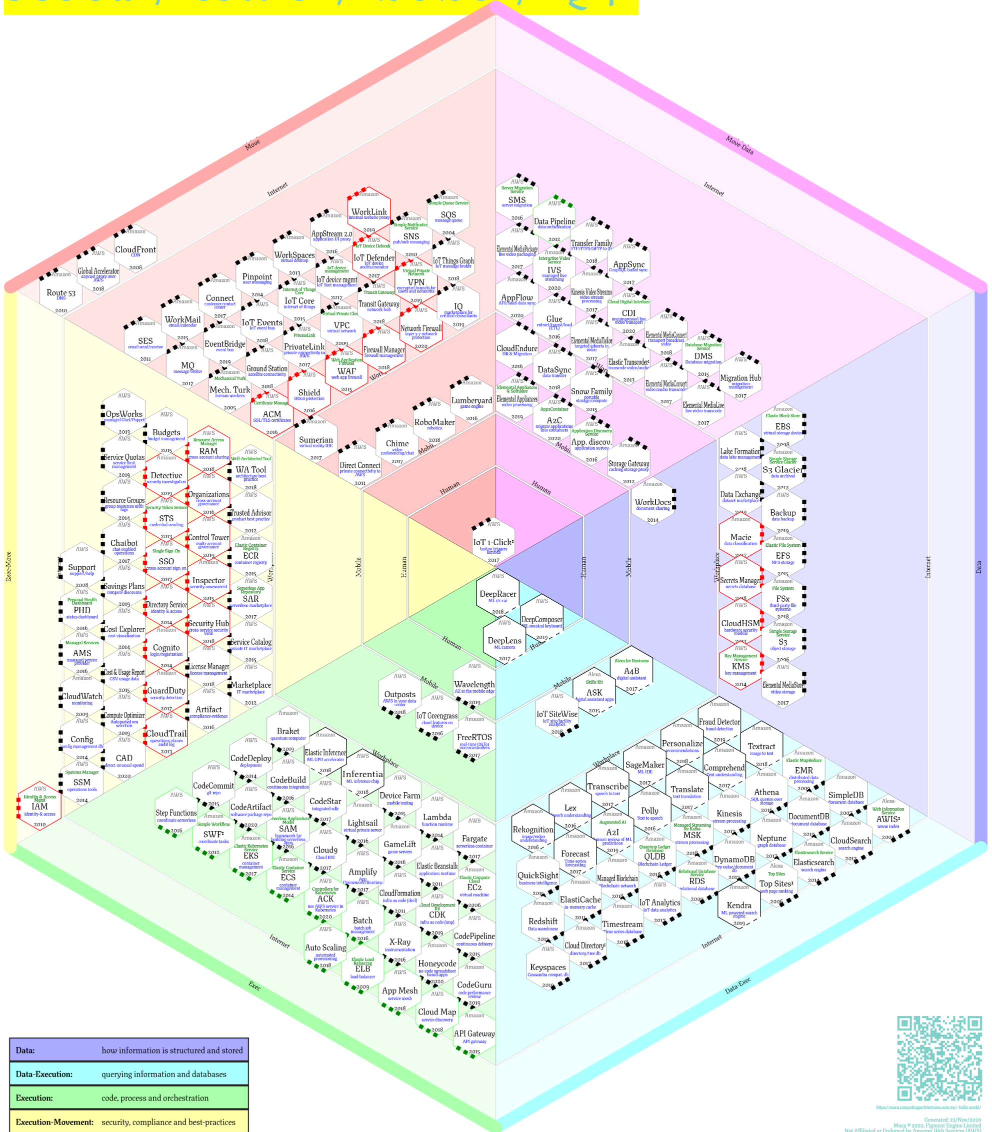




Quarter	Q1		Q2	
Year	Revenue	YTD Revenue	Revenue	YTD Revenue
2015	$45,186	$45,186	$70.609	$115,795
2016	$52,154	$52154	$73,542	$125,696
2017	$51,388	$51,388	$68,149	$118,537
2018	$48,281	$48,281	$66,853	$115,134
2019	$53,145	$53,145	$49,135	$102,280

Data warehousing using AWS services

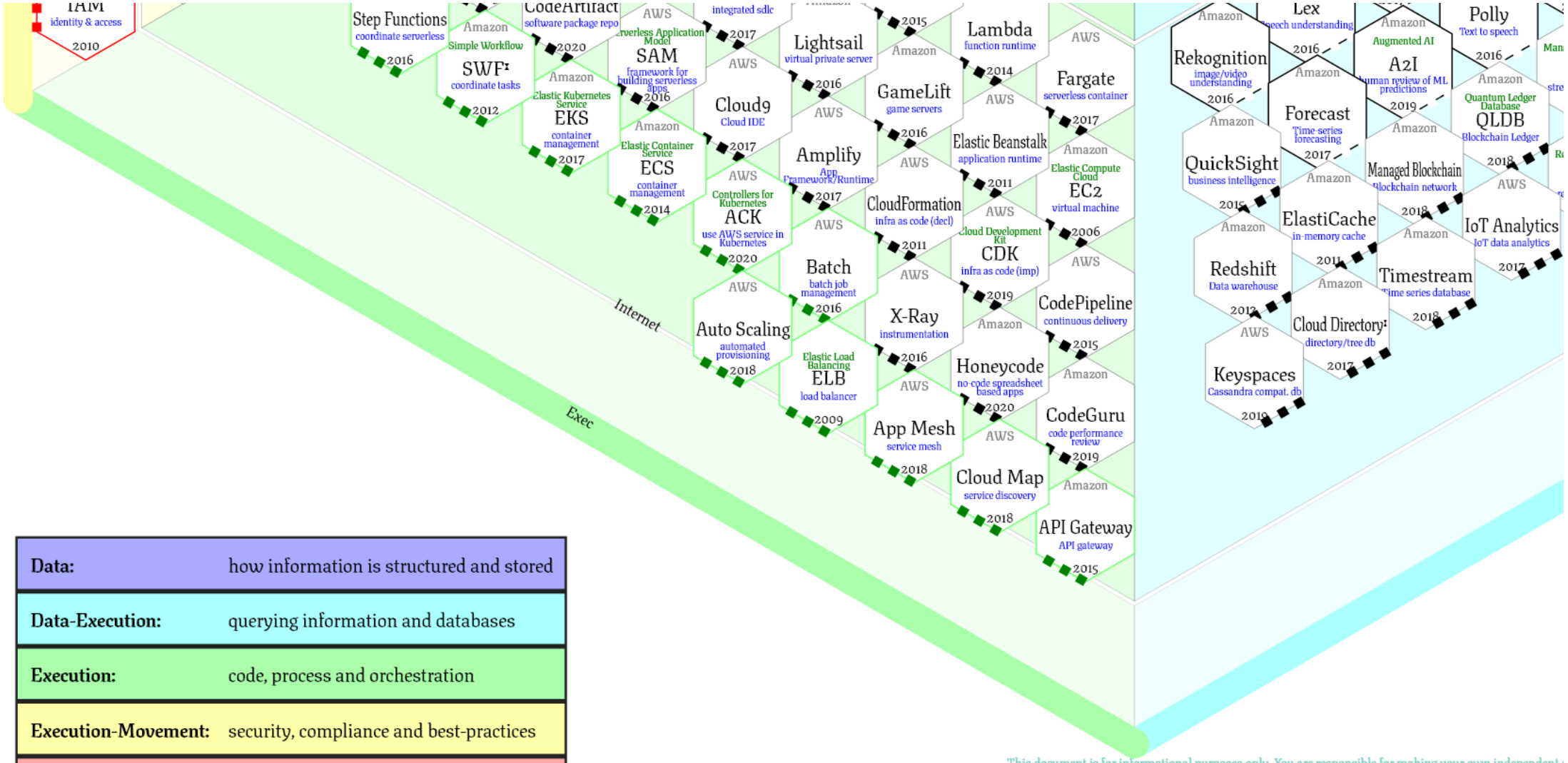
Moca / AWS / 2020 / Q4



Data:	how information is structured and stored
Data-Execution:	querying information and databases
Execution:	code, process and orchestration
Execution-Movement:	security, compliance and best-practices
Movement:	communication and messaging
Movement-Data:	data transfer and migration

Figment Engine Limited

Philip Fitzsimons



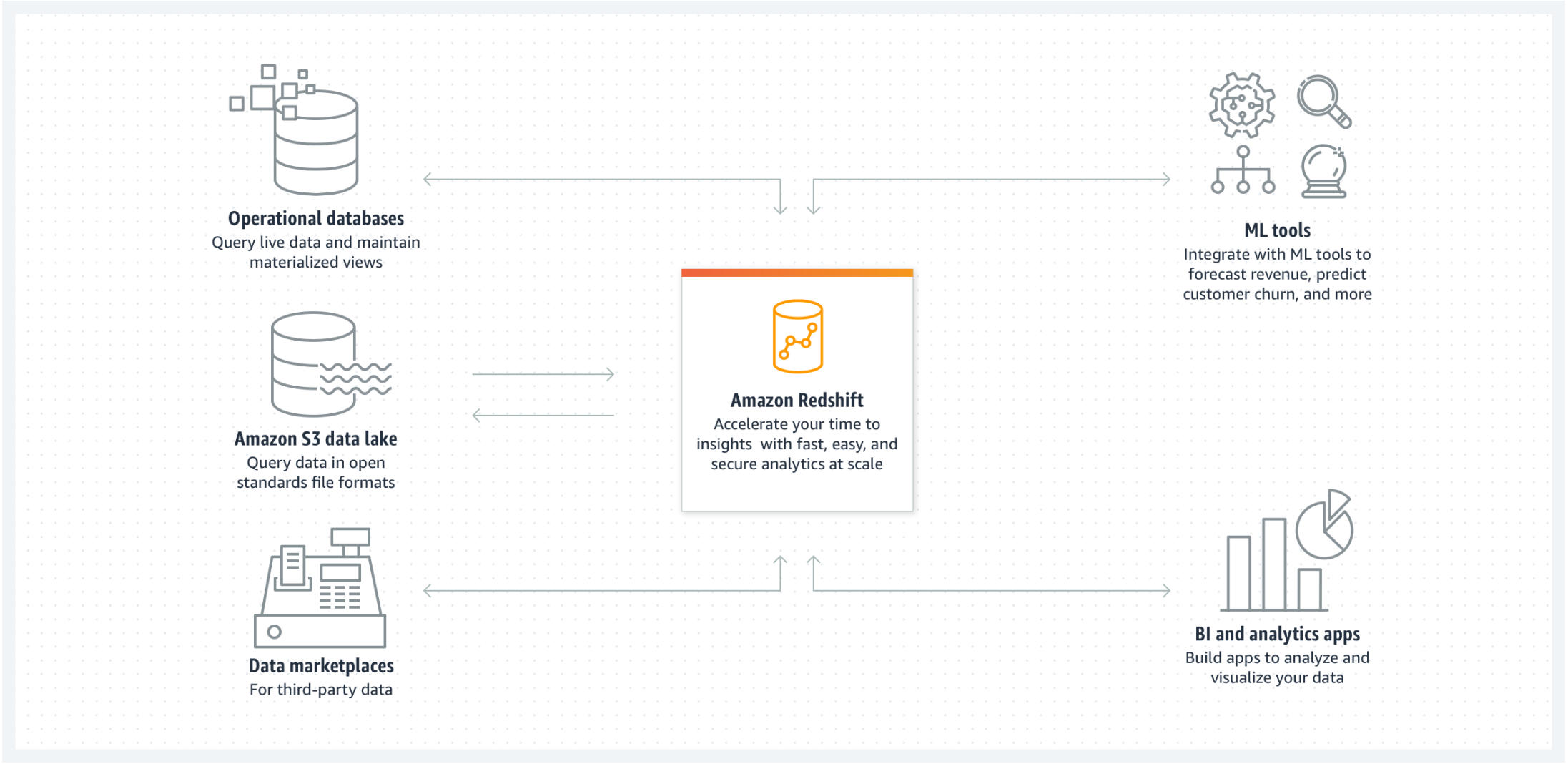
Data:	how information is structured and stored
Data-Execution:	querying information and databases
Execution:	code, process and orchestration
Execution-Movement:	security, compliance and best-practices
Movement:	communication and messaging
Movement-Data:	data transfer and migration

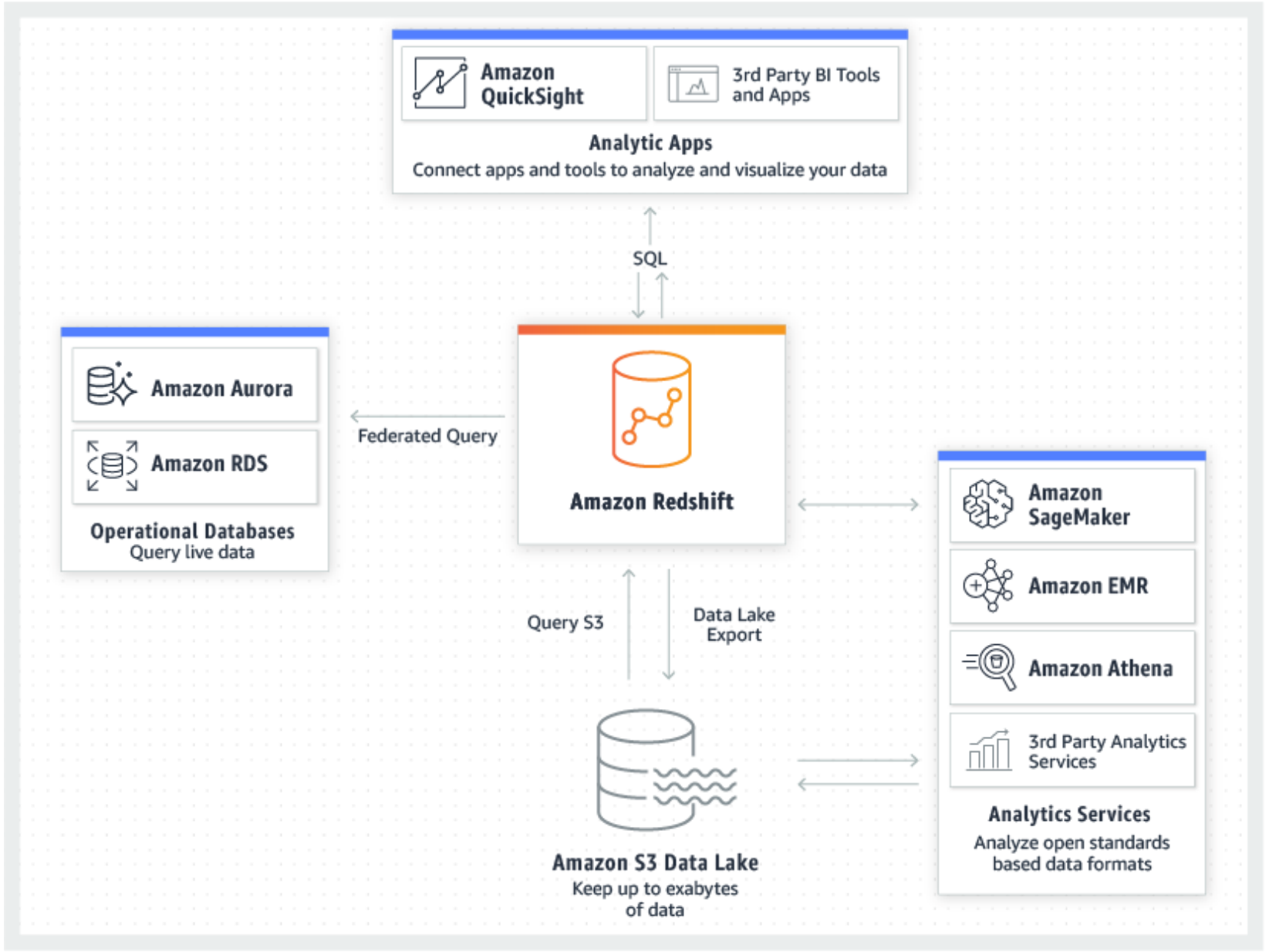
5 060780 440002

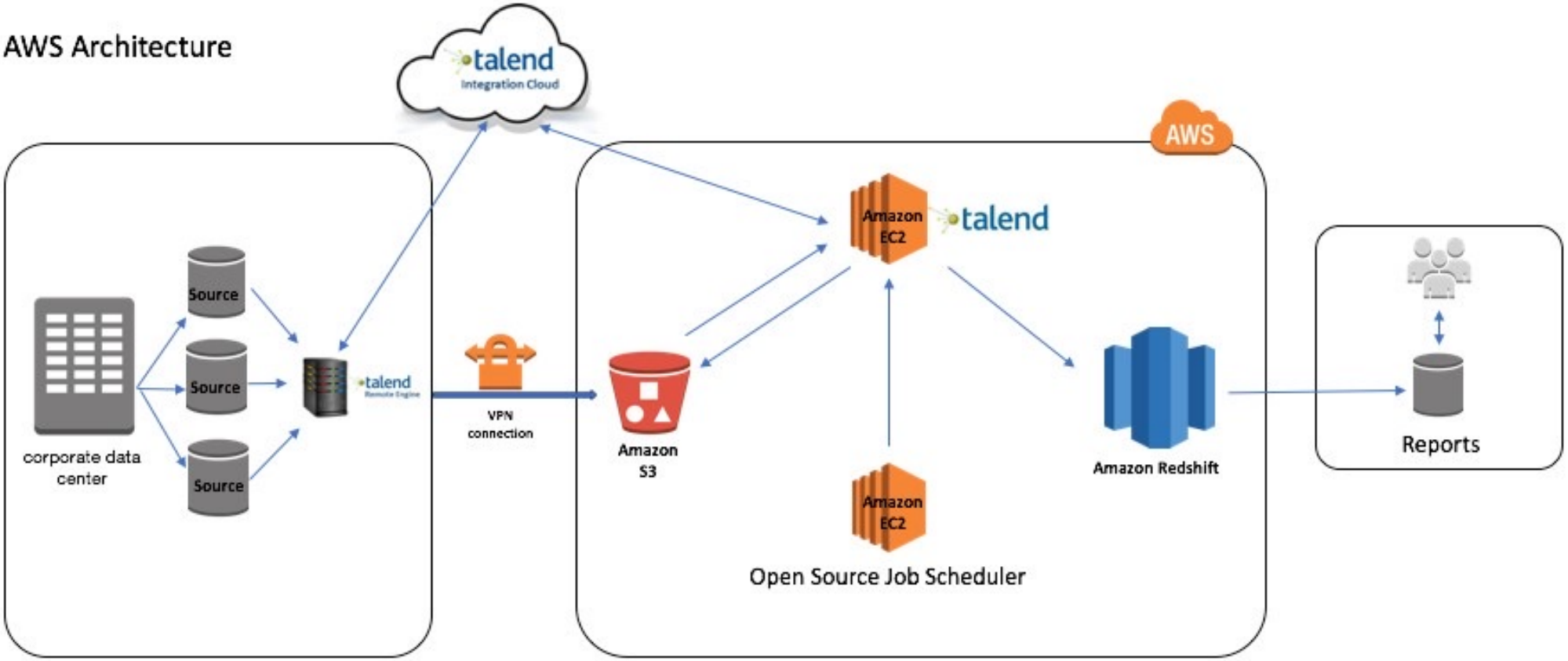
Figment Engi

Amazon RedShift

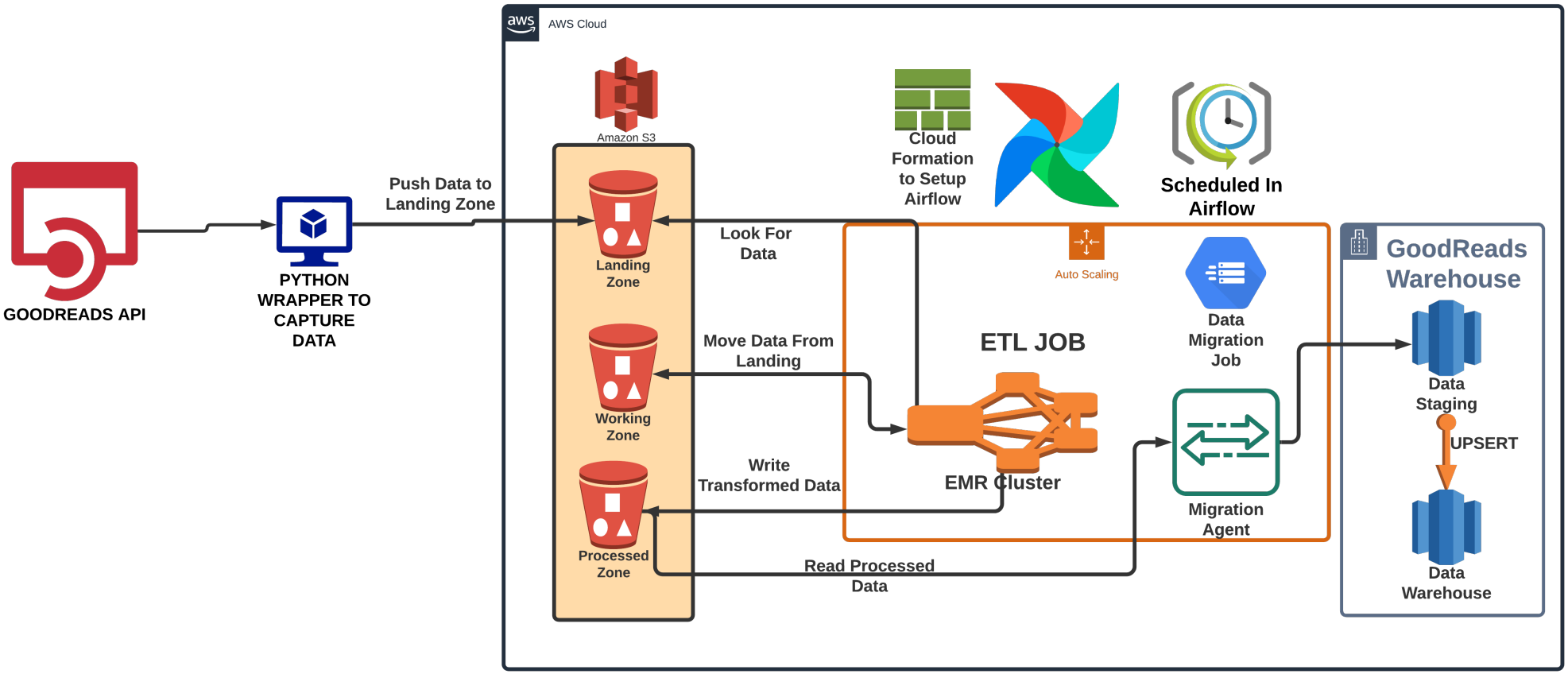
- Enterprise-class data warehouse and relational database query and management system
- Connect using many types of client applications
  - Business Intelligence (BI)
  - Reporting
  - Analytics
- Build multi-stage query operations that retrieve, compare, and evaluate large amounts of data
- Optimized storage and query performance
  - Massively parallel processing
  - Columnar data storage
  - Data compression encoding schemes


Operational databases
Query live data and maintain materialized views
Amazon S3 data lake
Query data in open standards file formats
Data marketplaces
For third-party data
Amazon Redshift
Accelerate your time to insights with fast, easy, and secure analytics at scale
ML tools
Integrate with ML tools to forecast revenue, predict customer churn, and more
BI and analytics apps
Build apps to analyze and visualize your data


Amazon QuickSight
3rd Party BI Tools and Apps
Analytic Apps
Connect apps and tools to analyze and visualize your data
SQL
Amazon Aurora
Amazon RDS
Operational Databases
Query live data
Federated Query
Amazon Redshift
Amazon SageMaker
Amazon EMR
Amazon Athena
3rd Party Analytics Services
Analytics Services
Analyze open standards based data formats
Query S3
Data Lake Export
Amazon S3 Data Lake
Keep up to exabytes of data



https://aws.amazon.com/blogs/database/using-amazon-redshift-for-fast-analytical-reports/



https://github.com/san089/goodreads_etl_pipeline