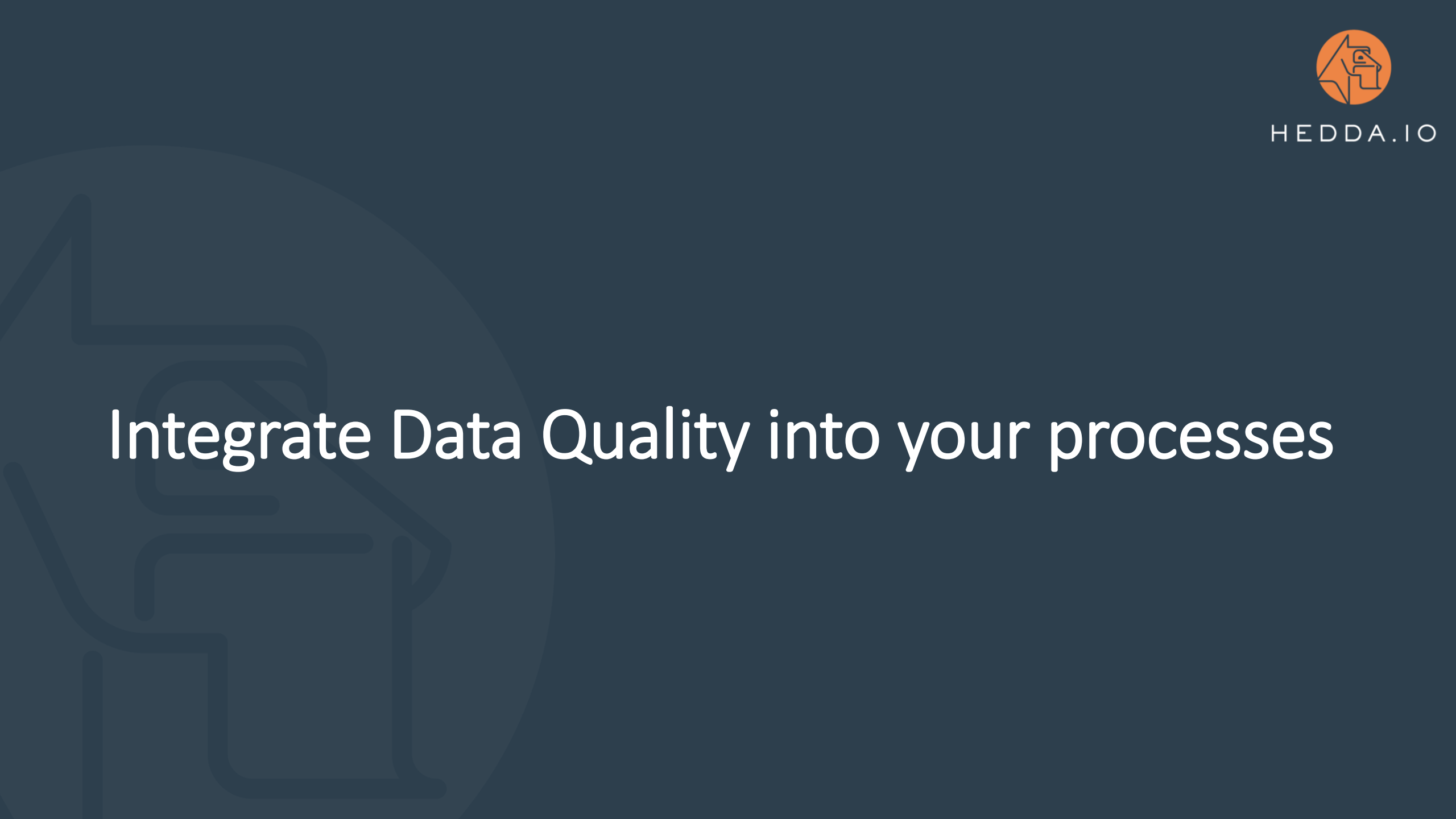


HEDDA.IO


HEDDA.IO
Integrate Data Quality into your processes

We are…

HEDDA.IO



Tillmann Eitelberg
CEO oh22information services GmbH
Microsoft Data Platform MVP



Rafael Dabrowski
oh22data AG
Lead Application Architect

What is Data Quality?

HEDDA.IO

- Data quality is never defined equal
- Each stakeholder has their own definition of the quality of data depending on their use case and the nature of the data they are handling
- Data quality can change over time, information can simply become outdated, or information requirements change
- The definition of data quality is often referred to as "fit for use"

The problem of data quality



- Data quality issues often look very simply, but a consistent and recurring cleanup that covers all potential cases is difficult to implement
- Data quality usually gets worse the more systems are connected
- Many data quality problems are often not immediately apparent
- Data quality issues are mostly solved by Data Engineers / Data Scientists and have no business context
- Most companies have not prioritized the topic of data quality or do not have it on their agenda

Typical Data Quality Problems



- Incorrect or outdated information (e.g., wrong customer addresses)
- Information that is no longer relevant or current (e.g., expired customer contracts or obsolete product specifications)
- Missing fields or records (e.g., no email provided in a customer database)
- Conflicting values across systems (e.g., different product prices in sales and inventory databases)
- Data not matching required formats or rules (e.g., invalid date formats)
- Data stored in multiple formats that make integration difficult (e.g., dates written as MM/DD/YYYY vs. DD.MM.YYYY)

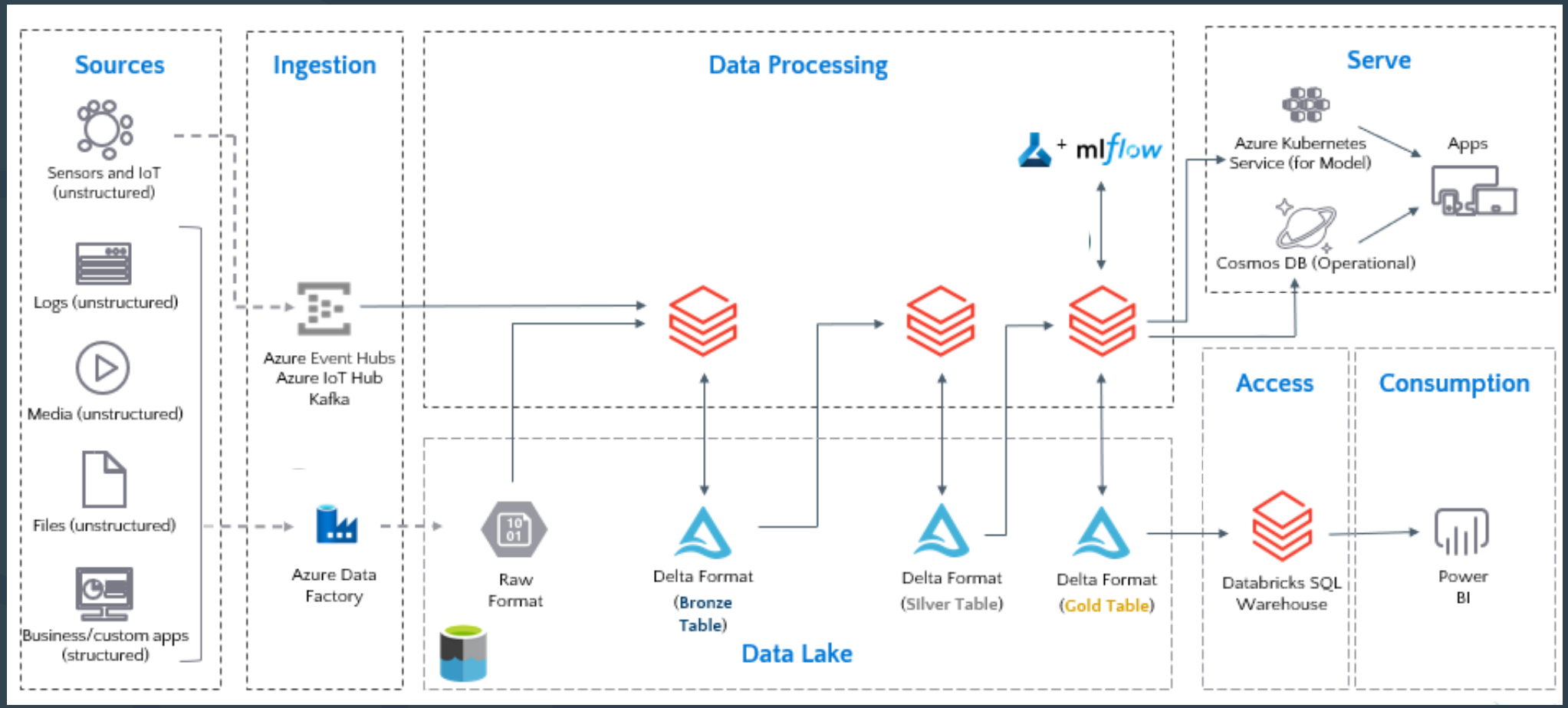
Typical Data Quality Problems



- Multiple entries for the same entity (e.g., duplicate product IDs)
- Lack of uniform data conventions (e.g., inconsistent abbreviations for countries: "US" vs. "USA")
- Errors during data summarization or grouping (e.g., miscalculating totals due to duplicate entries)
- Data that is cut off due to storage or field-length limitations (e.g., names or addresses missing important parts)
- Records that lose their relationship to parent entities (e.g., an order without a corresponding customer)
- Conflicting timestamps or records out of sequence (e.g., a transaction timestamped before its creation)

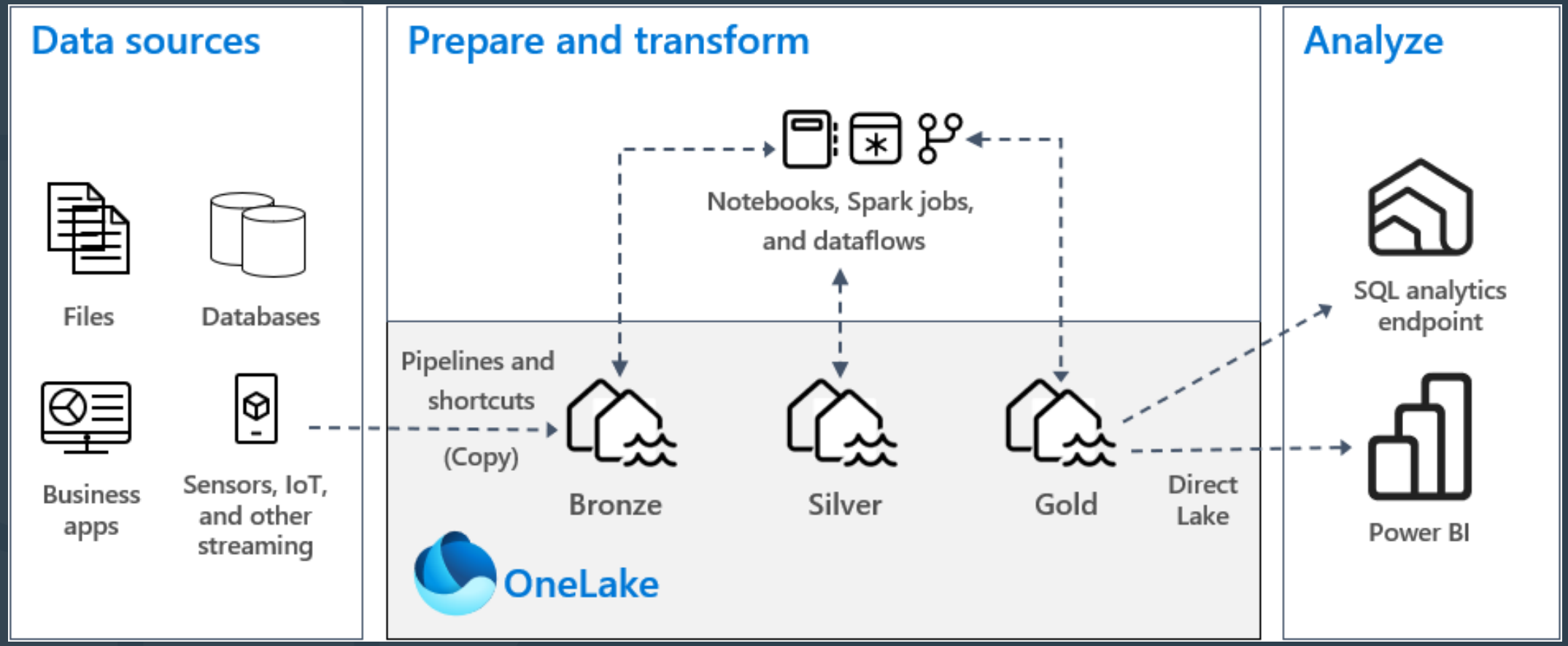
Data Quality Processes in a Lakehouse

HEDDA.IO



Data Quality Processes in a Lakehouse

HEDDA.IO



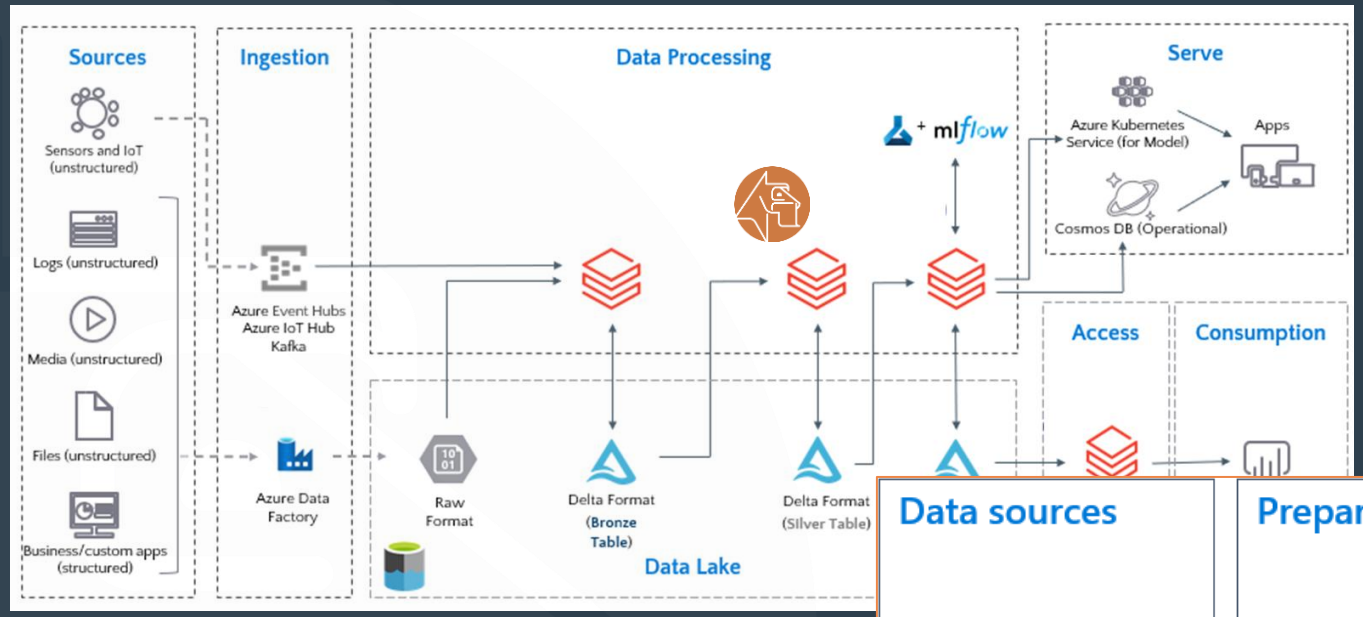
Data Quality Processes in a Lakehouse

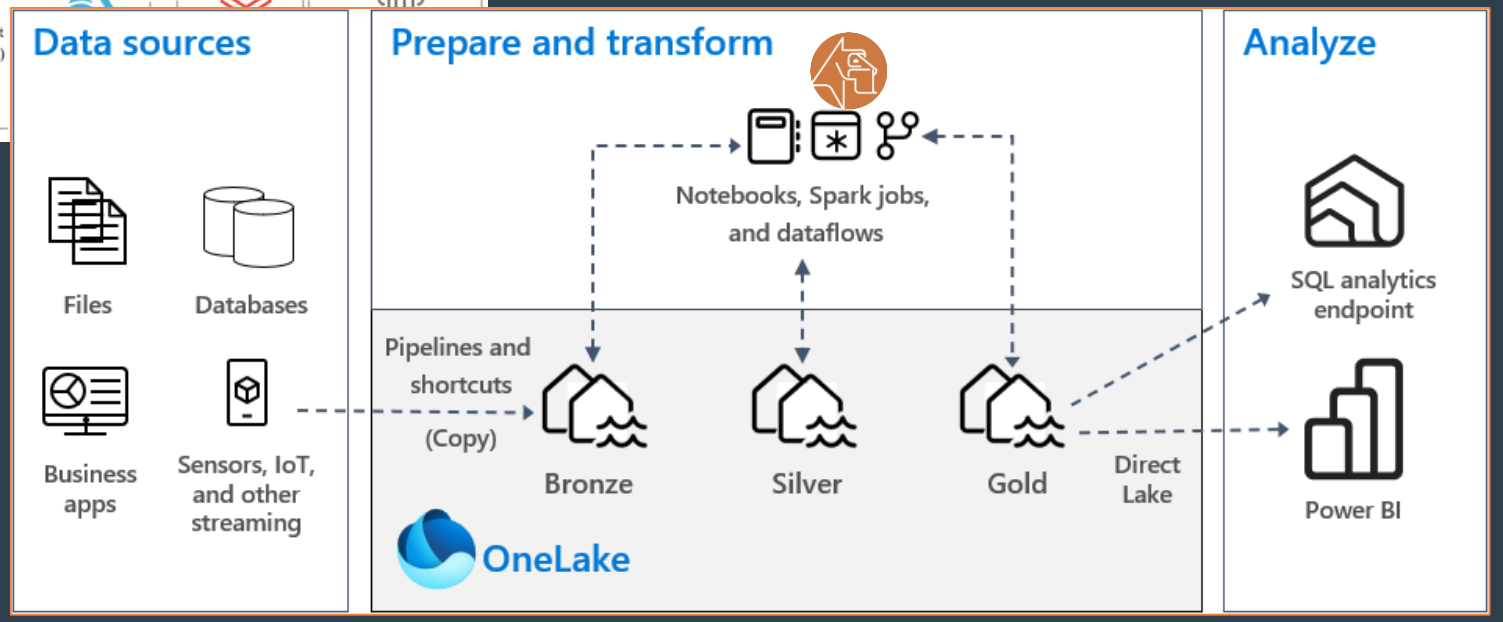


- Fill in missing data using methods like mean/median imputation, forward filling, or referencing external sources
- Remove rows or columns with excessive missing data
- Standardize inconsistent formats (e.g., date formats like MM/DD/YYYY vs. DD-MM-YYYY) and units (e.g., pounds vs. kilograms)
- Fix incorrect entries through cross-referencing with trusted sources (e.g., correcting misspelled names or wrong IDs)
- Detect and address outliers using statistical methods or business rules (e.g., flagging extreme values as errors)
- Uniformly format data, such as ensuring consistent casing (e.g., all-uppercase or all-lowercase) or aligning address formats
- Add missing context by integrating external data source
- Parse and split specific fields (e.g. product names, packaging units)

Data Quality Processes in a Lakehouse

HEDDA.IO





Data Quality Processes



- Conduct checks directly at the source system to ensure data consistency and correctness before ingestion
- Check incoming data for a valid schema, null values and other relevant information before further processes start
- Monitor and validate Sensor/IoT data directly as it arrives in the streaming layer
- Detect anomalies such as missing data, unusual patterns, or out-of-range values in real time
- Quarantine problematic records for manual review or automated remediation
- Verify joins and filters to ensure no rows are lost or misaligned

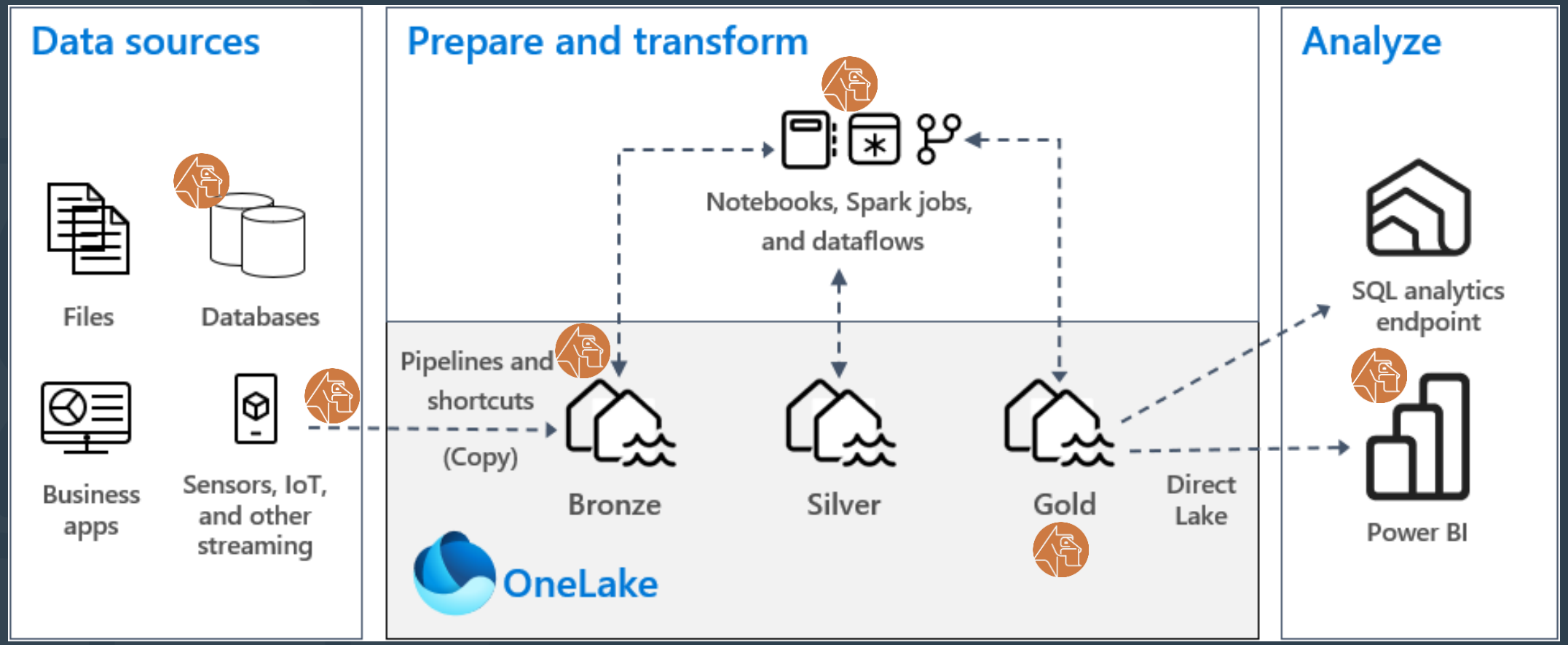
Data Quality Processes



- Validate the accuracy of transformations, such as field mappings and type conversions
- Test aggregated metrics like revenue or customer counts for accuracy
- Ensure accuracy and correctness of measures and DAX expressions in reports and dashboards
- Perform additional data cleansing and validation specific to machine learning requirements
- Address issues like outliers, class imbalance, and incomplete records to prepare high-quality training datasets
- Set up alerts for anomalies, such as unexpected increases in processing time or invalid rows

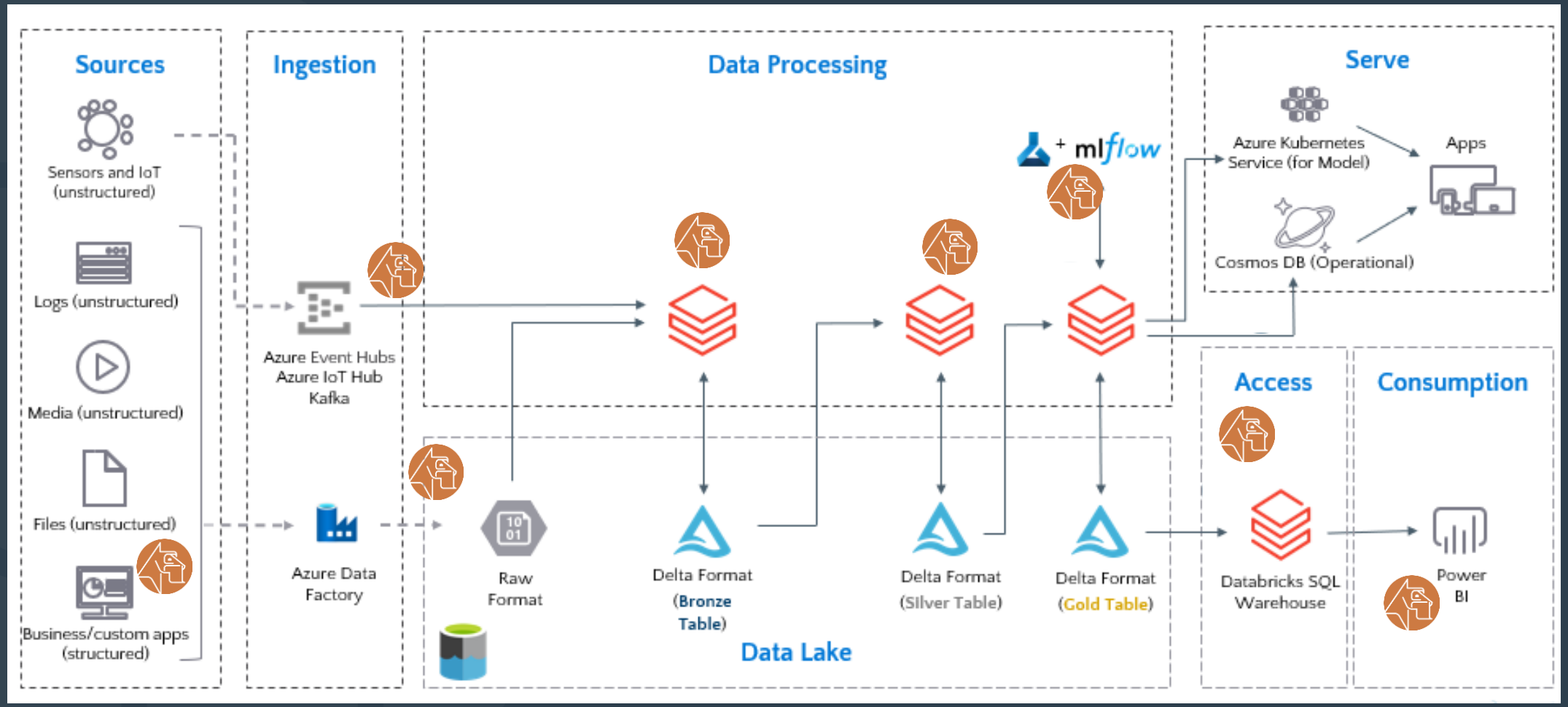
Data Quality Processes

HEDDA.IO



Data Quality Processes

HEDDA.IO



HEDDA.IO

HEDDA.IO

Let's start with HEDDA.IO

