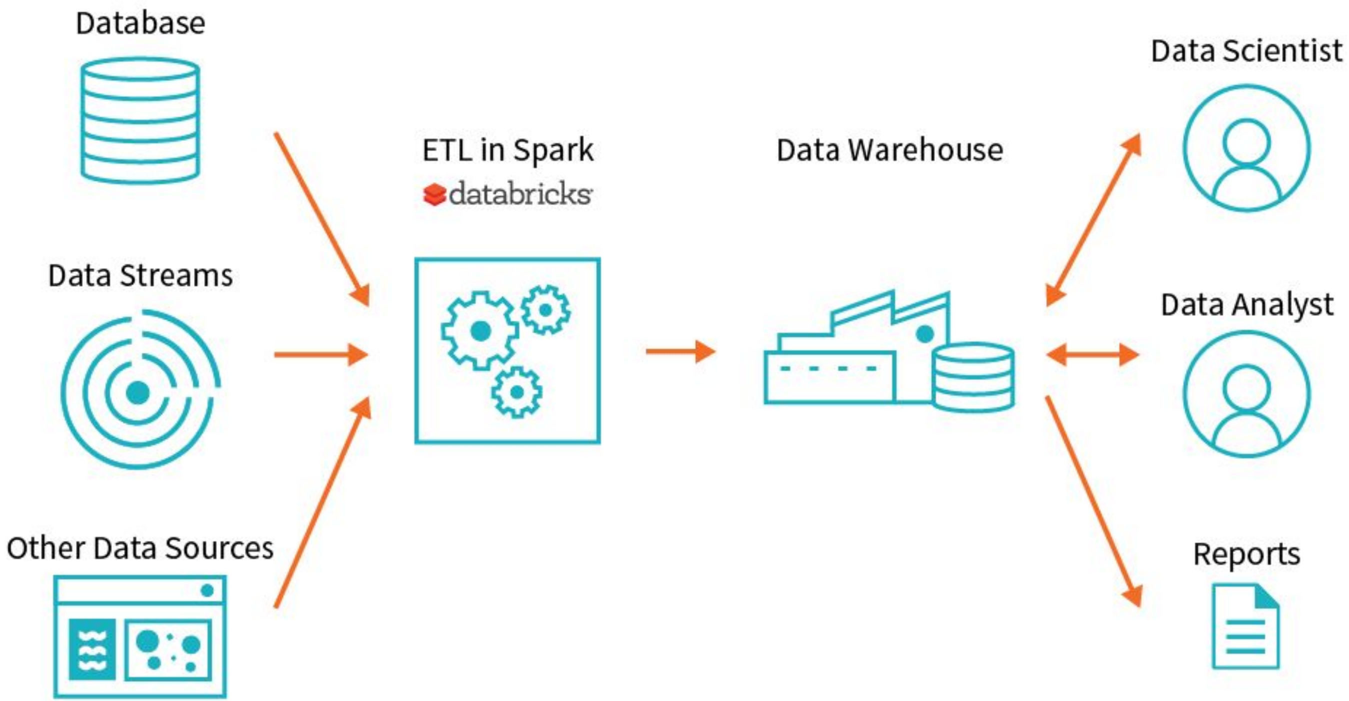
ETL 1

Estudos certificação Databricks

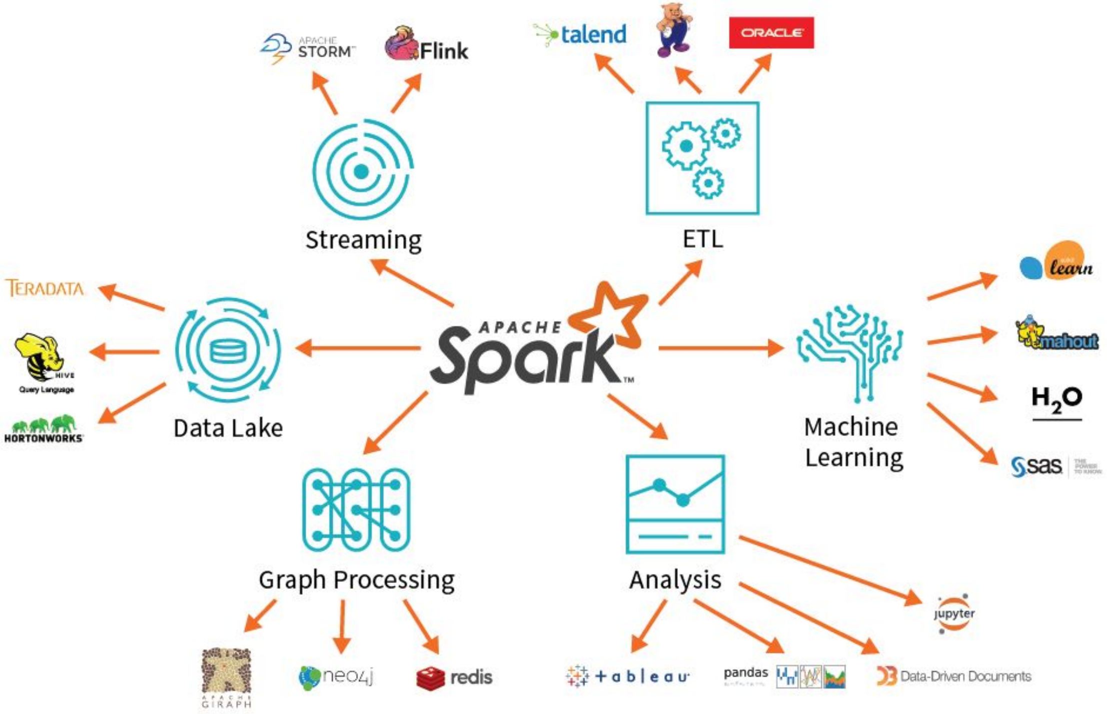
FLOW


Database
Data Streams
Other Data Sources
ETL in Spark
databricks
Data Warehouse
Data Scientist
Data Analyst
Reports

CONCEPTS

- Optimizing data formats and connections
- Determining the ideal schema
- Handling corrupt records
- Automating workloads

THE SPARK APPROACH



Streaming

ETL

Data Lake

Machine Learning

Graph Processing

Analysis

Data validation

One aspect of ETL jobs is to validate that the data is what you expect. This includes:

- Approximately the expected number of records
- The expected fields are present
- No unexpected missing values

REVIEW

Question: What does ETL stand for and what are the stages of the process?

Answer: ETL stands for `extract-transform-load`

1. *Extract* refers to ingesting data. Spark easily connects to data in a number of different sources.
2. *Transform* refers to applying structure, parsing fields, cleaning data, and/or computing statistics.
3. *Load* refers to loading data to its final destination, usually a database or data warehouse.

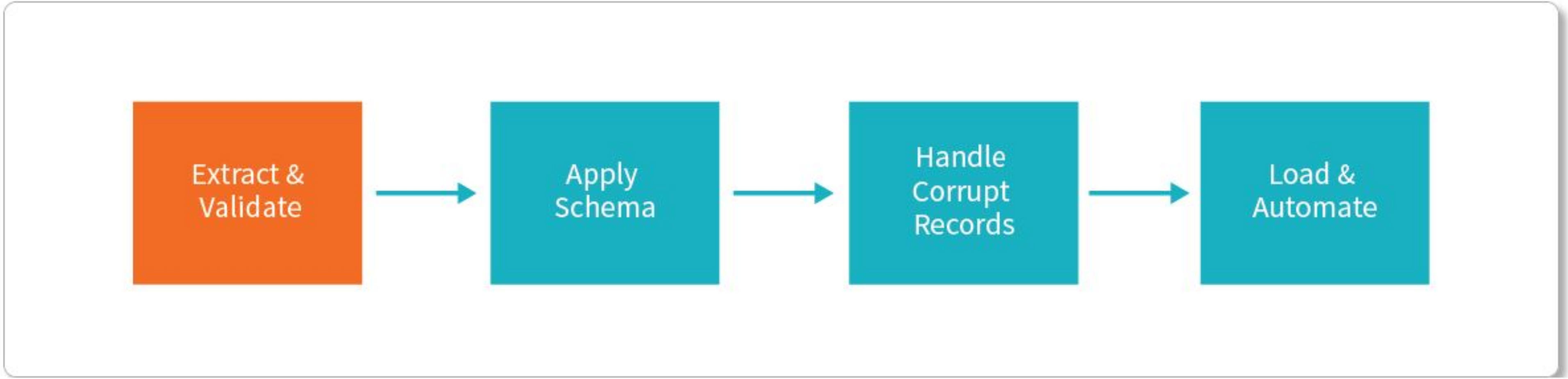
Question: How does the Spark approach to ETL deal with devops issues such as updating a software version?

Answer: By decoupling storage and compute, updating your Spark version is as easy as spinning up a new cluster. Your old code will easily connect to S3, the Azure Blob, or other storage. This also avoids the challenge of keeping a cluster always running, such as with Hadoop clusters.

Question: How does the Spark approach to data applications differ from other solutions?

Answer: Spark offers a unified solution to use cases that would otherwise need individual tools. For instance, Spark combines machine learning, ETL, stream processing, and a number of other solutions all with one technology.

STEP 1


Extract & Validate

Apply Schema

Handle Corrupt Records

Load & Automate

SETUP

Define your Azure Blob credentials. You need the following elements:

- Storage account name
- Container name
- Mount point (how the mount will appear in DBFS)
- Shared Access Signature (SAS) key

https://docs.databricks.com/data/data-sources/azure/azure-storage.html#mount-azure-blob-storage-containers-with-dbfs

DATABASE CONECTION

```
jdbcHostname = "server1.databricks.training"

jdbcPort = 5432

jdbcDatabase = "training"

jdbcUrl = f"jdbc:postgresql://{jdbcHostname}:{jdbcPort}/{jdbcDatabase}"
```

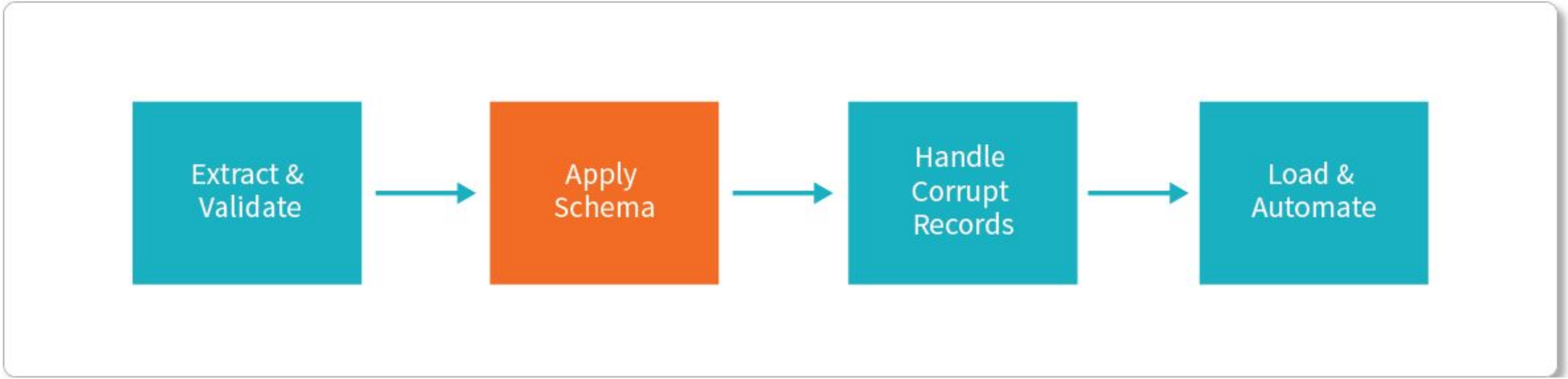
DATABASE CONECTION

```
connectionProps = {
  "user": "readonly",
  "password": "readonly"
}
```

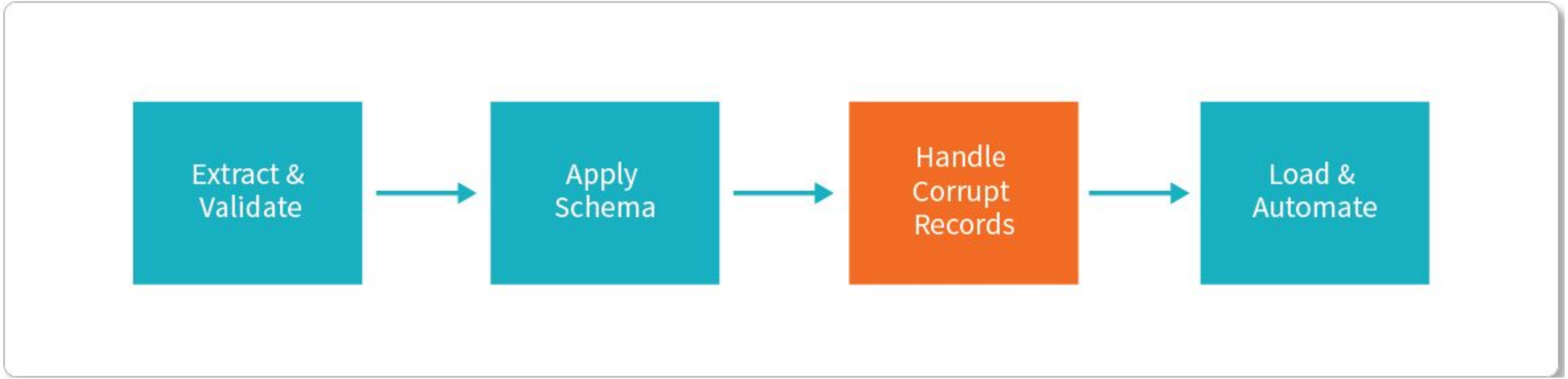
READ THE DATABASE

```
tableName = "training.people_1m"


peopleDF = spark.read.jdbc(url=jdbcUrl, table=tableName,
properties=connectionProps)


display(peopleDF)
```

STEP 2


Extract & Validate
Apply Schema
Handle Corrupt Records
Load & Automate

STEP 3



Extract & Validate

Apply Schema

Handle Corrupt Records

Load & Automate

STEP 4