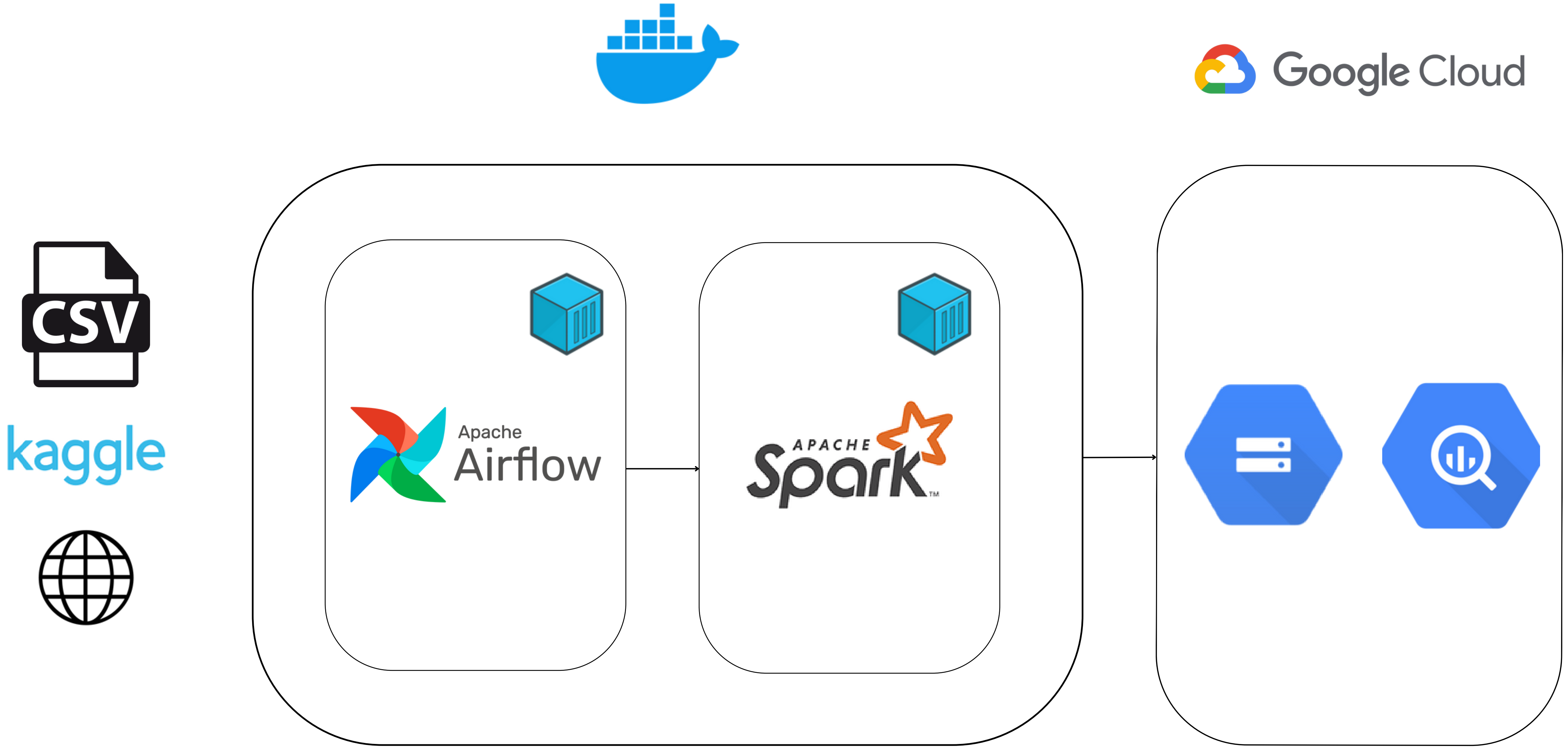
CSV
kaggle
Apache Airflow
APACHE Spark
Google Cloud

Architecture overview

Components

- Data Source
- Data Ingestion
- Data Processing
- Data Analytics and Querying
- etc.

Guideline

- Version for each component
- Signs and Symbols
- Description of the process on the arrow

Flow Diagram

“contains the diagram regarding how the data pipeline works, where the data is being consumed and produced, from the source to the destination in the form of flow charts”

DAG Document

```
## 📝 DAG Documentation - {job_name}
> <small>{Description}</small>

#### ⚙️ Default Arguments
&ensp;&ensp; 🧑‍💻 **Owner**: &ensp; `{owner}` &emsp; | &emsp; 🕒 **Schedule**: &ensp; `{schedule}` &emsp; | &emsp; 📅 **Start Date**: &ensp; `{start_date}` &emsp; | &emsp;

#### 📋 Pipeline Info
- 📌 **Source**: &ensp; `{source}`
- 🗂️ **Source Data**: &ensp; `{source_data}`
- 📦 **Destination**: &ensp; `{destination_data}`
- 🔗 **Github Link**: &ensp; [{job_name}]({job_repo})

#### 📞 Contact
&ensp;&ensp; 📧 **Requestor Team**: &ensp; `{requestor_team}` &emsp; | &emsp; 👥 **Source Team**: &ensp; `{source_team}` &emsp; | &emsp; 🧑‍💻 **Users Team**: &ensp; `{user_team}`
```

Exploratory Data Analysis

The process of investigating and analyzing data sets to summarize their main characteristics, often using visual methods. For data engineers, EDA is a critical step when working with raw data to ensure it is clean, consistent, and well-understood before building pipelines, performing transformations, or making it available for data analysis or machine learning tasks.

Checklist

1. Data Overview - check dimension column name and column type
2. Preview Data
3. Missing Value
4. Duplicates
5. Outliers
6. Numeric data or Categorical data
7. Distribution of data
8. Correlations of data - correlations method or heatmap
9. Grouping data - by categories or aggregate metrics

Cleaning Data and Validation

Cleaning Data

- Missing Data
- Data Transformation - Normalize data and Standardize format
- Correct Anomalies - fix incorrect data and replace data

Validate Data

- Schema Validation - ensure that column names and data types match expected schemas.
- Data integrity Checks - Verify constraints (e.g., unique keys, foreign key references) and Check for unexpected nulls or deviations from expected patterns .