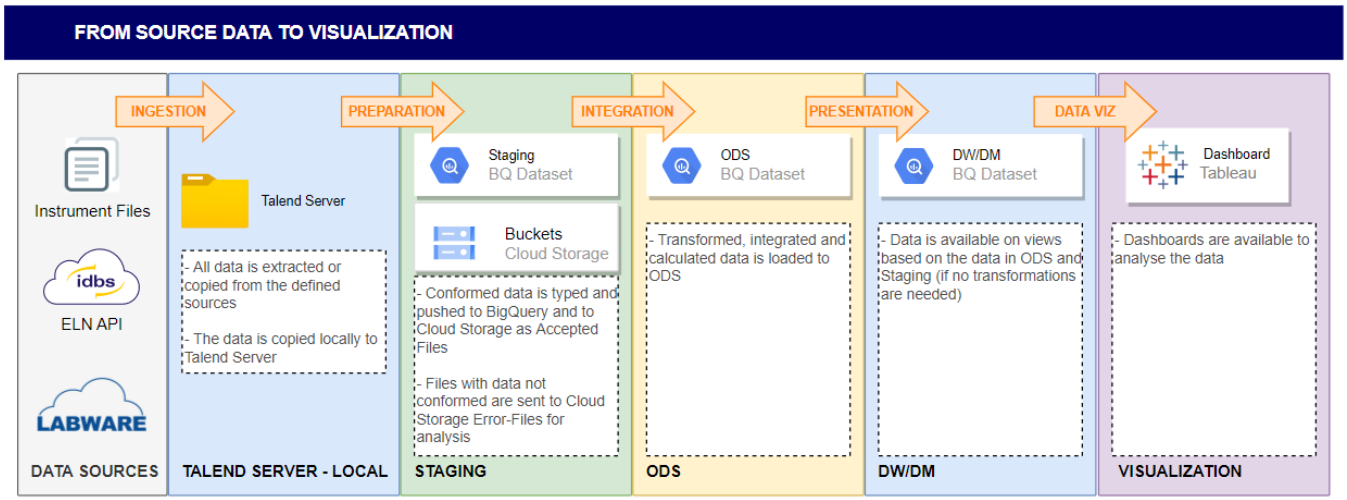
# ALB Data Dev Architecture - General

This page presents the general architecture for data development. It's responsable to describe the data extraction, ingestion, preparation, integration and share. Not all the domains are using this architecture. Please, refers to the other subpages of **Data Development** to get to know the details and the differences.

## Summary



## Macro Diagram - ETL Phases




## 1 - Data Ingestion

Data ingestion is the process of transporting data from one or more sources to a target site for further processing and analysis. For this project data is extracted or copied to Talend server as files using Talend.

### Data Sources

#### ELN

Many spreadsheets are extracted from Electronic Laboratory Notebook (ELN) which is responsable to gather lab documents and result tests. It gives scientists a common view of data across disparate research areas, enabling complete visibility of research information.

ELN data is extracted in JSON format using an available API.

Related documents:

ZIFO - API Documentation for extracting E-Workbook Spreadsheet Data

#### Labware

LabWare LIMS is a laboratory information management system (LIMS) that is "**configurable or pre-configured for laboratories of all types and sizes**." The software consists of the core LIMS application with access to LabWare's library of LIMS software modules.

#### Instruments/Cyclers

The instruments files are available in different formats (txt, csv, xml, etc.) usually on Lab server folders or Google drive. The rules to structure the files in columns are specific for every workflow/method.

# 2 - Data Preparation

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. Often involves reformatting data and making corrections to data.

Talend jobs are responsable for cleaning and structuring the data.

In this process, we also define if a source file should go to **Accepted Files** or **Rejected Files** bucket. The files with the expected structure go to Accepted and the files that we cannot load on BigQuery for a structure or data type issue go to Rejected.  The process will not stop in case of rejected files.

# 3 - Data Integration and Enrich

Data enrichment refers to the process of appending or otherwise enhancing collected data with relevant context obtained from additional sources.

# 4 - Data Presentation

All the reports, dashboards and users should access the data using this layer. Data marts (DM) with views will be created by subject, limiting access to non treated data.

# 5- Data Visualization

Check with the Data Viz team the specific documentation.