Lesson 4

Data Sources, Data Quality, Preprocessing and Data Store Export to Cloud Services

Data Sources for the Applications, Programs and Analytics Tools

- Can be external, such as sensors, trackers, web logs, computer systems logs and feeds
- Can be machines, which source data from data-creating programs.
- Data sources can be structured, semi-structured, multi-structured or unstructured.

Data Sources for the Applications, Programs and Analytics Tools

- Data sources can be social media (L4 Data Processing Layer)
- A source can be internal. Sources can be data repositories, such as database, relational database, flat file, spreadsheet, mail server, web server, directory services

Data Sources

- Can be text or files such as comma-separated values (CSV) files
- Source may be a data store for applications (L4 Data Processing Layer)

Structured Data Sources

- SQL Server, MySQL, Microsoft Access database, Oracle DBMS, IBM DB2, Informix, Amazon SimpleDB or a file-collection directory at a server
- Data dictionary which enables references for accesses to data─ consists of a set of master lookup tables..

Unstructured Data Sources

- Distributed data over high-speed networks needing high velocity processing
- Sources are from distributed file systems. The sources are of file types, such as .txt (text file), .csv (comma separated values file)

Unstructured Data Sources

- Data may be as key-value pairs, such as hash key-values pairs
- Data may have internal structures, such as in e-mail, Facebook pages, twitter messages etc.
- The data do not model, reveal relationships, hierarchy relationships or object-oriented features, such as extensibility.

Data Quality

- Data quality is high when it represents the real-world construct to which references are taken
- High quality means data, which enables all the required operations, analysis, decisions, planning and knowledge discovery correctly.

Data Quality

- A definition for high quality data, especially for artificial intelligence applications, can be “ data with five R’s as follows: Relevancy, recency, range, robustness and reliability.”
- Relevancy is of utmost importance.

Data Integrity

- Refers to the maintenance of consistency and accuracy in data over its usable life
- Software, which store, process, or retrieve the data, should maintain the integrity of data

Noise

- Noise─ One of the factors effecting data quality
- Noise ─ refers to data giving additional meaningless information besides true (actual/required) information

Outlier

- Outliers─ A factor which effects quality
- Refers to data, which appears to not belong to the dataset
- For example, data that is outside an expected range.

Outlier

- Actual outliers need to be removed from the dataset, else the result will be effected by a small or large amount
- Outlier, if real, not because of error, can be useful in detecting anomaly, .

Missing Values

- Missing value ─ A factor effecting data quality
- Implies the data not appearing in the data set.

Duplicate Values

- Duplicate value— A factor effecting data quality
- implies the same data appearing two or more times in a dataset

Data Pre-processing

- An important step at the ingestion layer L2 in Data Processing Architecture
- Must before running a Machine Learning (ML) algorithm and Analytics
- Needed before the Data exported to a data store or cloud service

Data Pre-processing Needs

(i) Dropping out of range, inconsistent and outlier values

(ii) Filtering unreliable, irrelevant and redundant information

(iii) Data cleaning, editing, reduction and/or wrangling

Data Pre-processing Needs

(iv) Data validation, transformation or transcoding

(v) ELT processing. [Extract, Load and Transform]

(vi) Enriching, Editing, Wrangling, Reduction

Data Formats for Data Transfer to Data Store or Cloud Services

- Transfer Formats from (a) data storage, (b) analytics application, (b) service or (d) cloud:

(i) CSV (Example 1.9), (ii) JSON (Example 3.3), (iii) Tag Length Value (TLV), (iv) Key-value pairs (Section 3.3.1), (v) Hash-key-value pairs (Example 3.2).

Figure 1.3 Data pre-processing, analysis, visualization, data store export

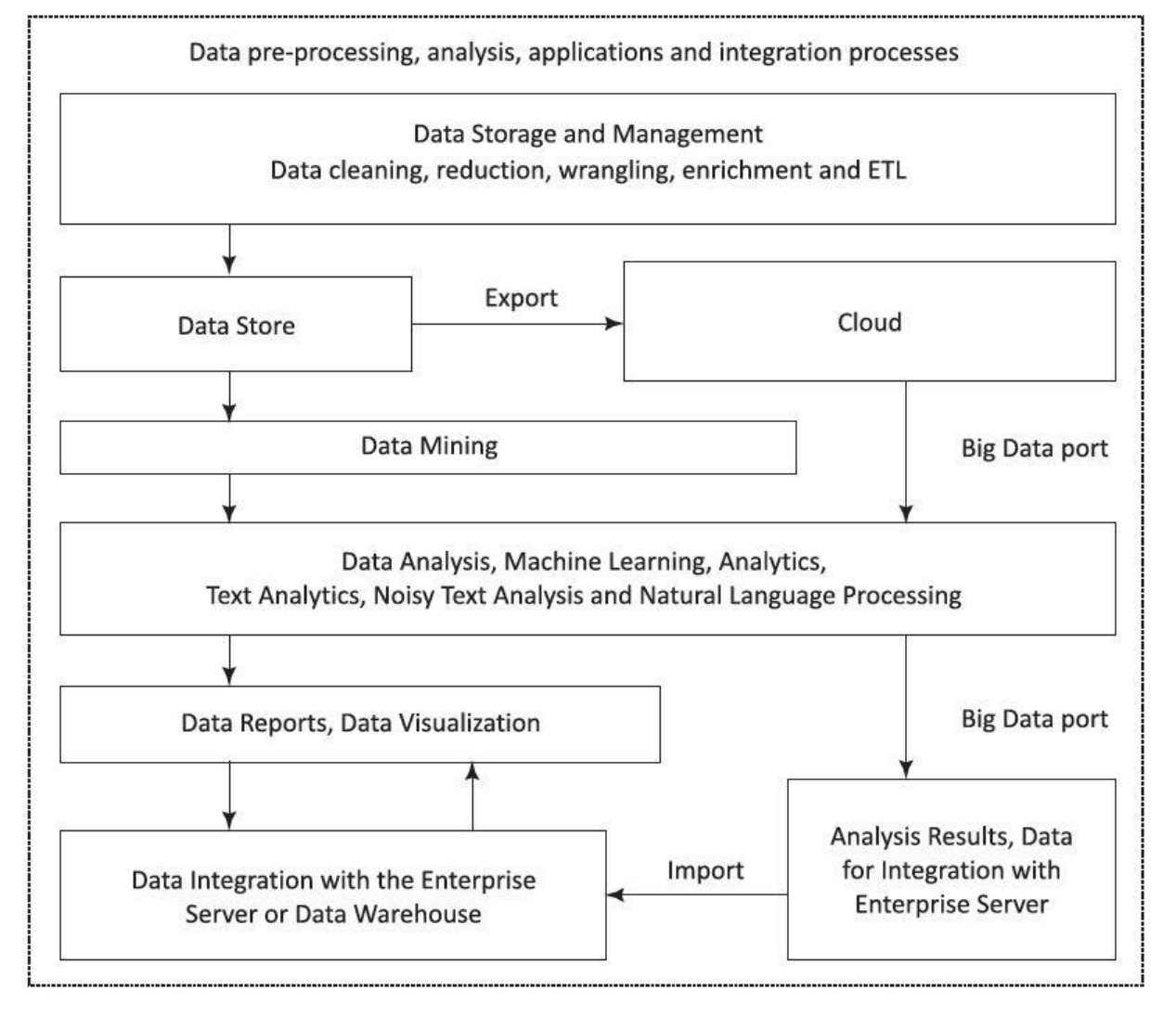


Figure 1.4 Data store export from machines, files, computers, web servers and web services

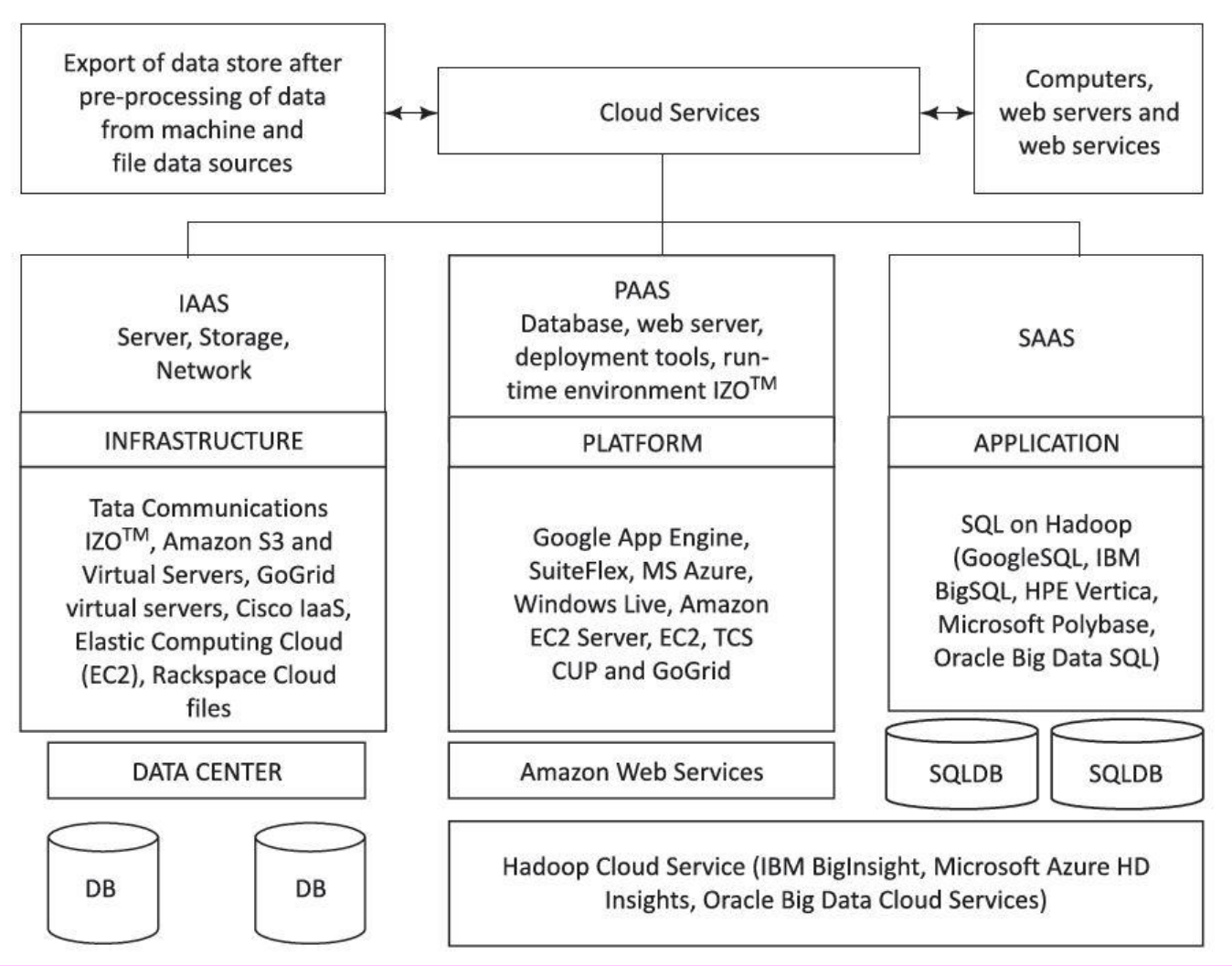
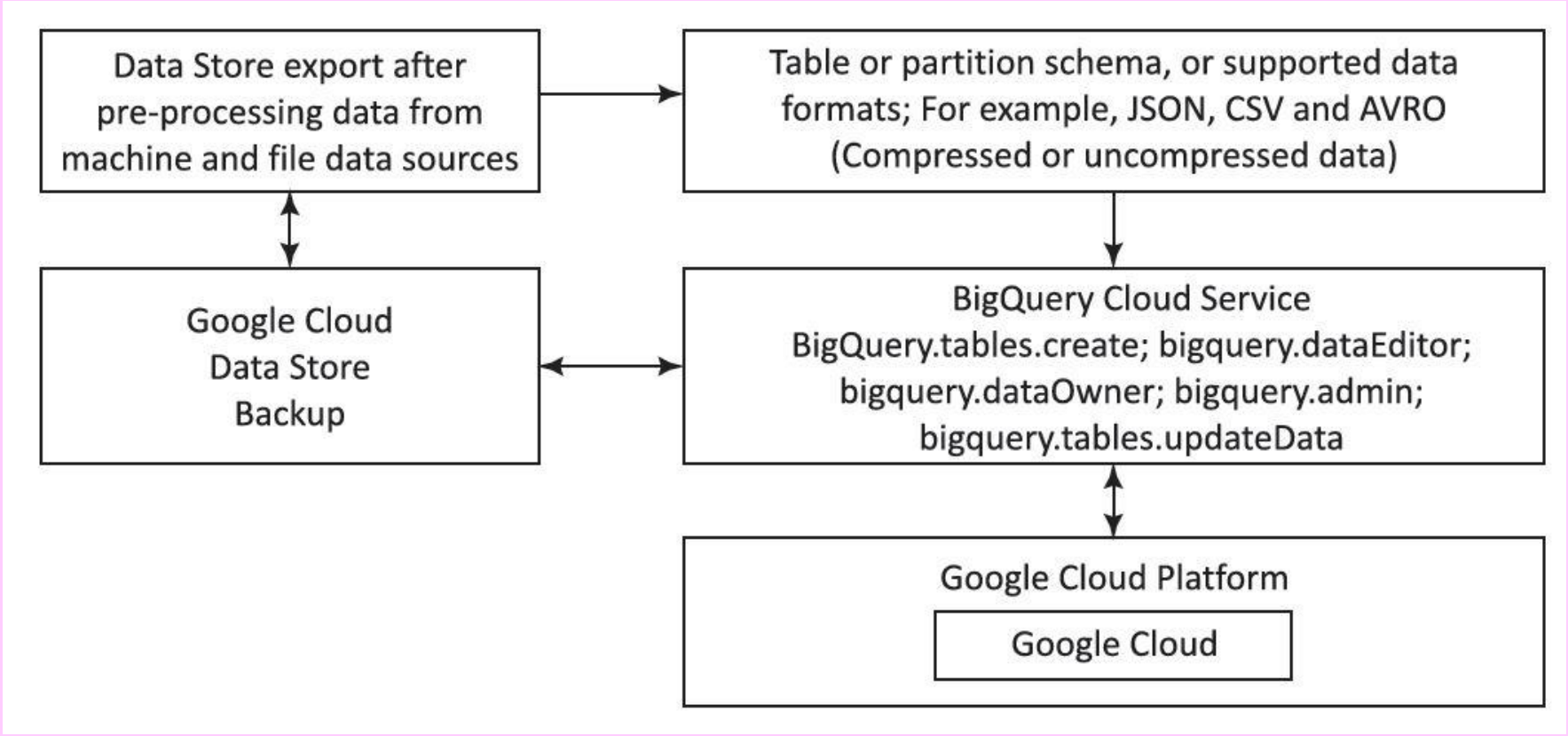


Figure 1.5 BigQuery cloud service at Google cloud platform



Summary

We learnt :

- Internal and External Data Sources
- Structured, Semi or unstructured
- Data Dictionary
- Can be text or files, CSV/JSON files
- Data store for application

Summary

We learnt :

- Data Quality, Integrity, Noise, Outliers, Mission Values
- Data Preprocessing Needs
- Cleaning, filtering
- Data store export from machines, files, computers, web servers and web services

End of Lesson 4 on

Data Sources, Data Quality, Preprocessing and Data Store Export to Cloud Services