Data Nuts & Bolts : Fundamentals of Data

Data – Raw & unstructured facts that convey a message

Information – contextualized, organized & vetted data that conveys some sort of trend or pattern

Data -> Information is one of the most important tasks of IT. The end info answers questions such as who, where, when, what & how

Knowledge – application of information. It’s measured by the ability to do things using the know how gained from Information.

Wisdom - Ability to use knowledge & experience to make good decisions

Sources of data generation & data formats

Common data collection types

Observational – Obtained from direct observation. Eg: Temp. of a room

Experimental - Obtained from research & investigation. Eg: DNA sequencing results

Compiled – Acquired through manual data gathering, aggregated over time. Eg: Data mining

Simulated – Results of studying certain behaviors & identifying new patterns. Eg:- Customer shopping behavior.

Common data generation sources

Human generated data – data from phones, emails, calls, web etc.

Machine generated data – IoT, medical equipment, satellites etc.

Business generated data – Banks, stock exchanges, etc.

Common data formats

Quantitative data – things that are measurable in terms of numbers & values

Qualitative data – things that are descriptive, generally not measurable, exploratory, subjective, usually unstructured

Geospatial data – has some kind of association with location

Digital data – images, audio, video etc.

Documentation & scripts – rich text format, open text document & HTML

Data formats w.r.t storage & processing

Structured – easiest to search & analyze, often stored in rows & cols

Unstructured – not contained in row-col database

Semi-structured – has some defining & consistent characteristics. It doesn’t conform to a structure as rigid as expected with a database. There are some organizational properties called semantic tags or metadata to make it easier to organize but there’s still fluidity in the data. Eg: XML, JSON

You can’t perform analytics on unstructured data. It has to be transformed to structed data.

Data Terminologies

Data can be classified into 2 groups:

Primary data – data a person collects by themselves

Secondary data – data collected by a third party

Database – repository for data

Data analytics – processes/technologies used to explore data

Data aggregation – procedures used to gather data

Metadata – details about a particular dataset. Data about data

Time series – analysis studying temporal patterns

Anonymized data – data in which individual’s identities are kept secret

Augmented reality – data analytics approach utilizing ML & NLP

Data literacy - ability to understand, read, create and communicate data as information

File format – structure used to identify & encode data

Data science – interdisciplinary field which uses scientific methods to extract knowledge & insight from data

Deep learning – neural network-based ML algorithms

Data projects are mainly categorized into 3 types:

Data pipelines & data staging

Data processing & analysis

Application development

Data storage & Backup

3 main methods of backup

Full backup –

Most comprehensive type of backup

Covers backup for all data within the hard drive

Requires high storage space

Slowest backup speed

Contains lots of duplicate data

Differential backup –

Getting a backup of the data that was generated/modified since the last FULL backup

Requires medium-high storage space

Fast backup speed

Contains some duplicate data

Incremental backup –

Storing only modifications that were made to the previous backup regardless of whether the previous backup was a full backup or any other kind of backup

Small storage space

Fastest backup speed

No duplicate data

Data migration & ETL

Data migration is one of the most complex & critical parts of every data transformation project.

Eg: Deciding to use a new CRM or ERP

Data migration involves 6 main steps:

Planning – Identify stakeholders, identify data, determine risk mitigation & backup strategies

Analyzing the data – explore the data, determine the impact of migration cut-off point, create a data dictionary

Design – develop “source to target mapping”, a dependency tree on order of migration is required if multiple data sets are to be migrated, determine ETL transformations required

Implementation – configure tools needed for migration, write migration scripts ensuring best practices

Testing – deploy all required tools, create a test plan with accurate data coverage, design a migration validation engine

Final migration – execute all data migration steps, ensure proper planning & most importantly get stakeholders to sign off on this plan, execute go-live including migration, define decommissioning plan for all systems that previously stored the data

What common data source formats are used in ETL process?

Relational databases, XML, JSON, flat files, may include non-relational database structures

In streaming ETL pipelines (a.k.a event-driven ETL), the extracted source data is loaded on-the-fly to the destination database without any intermediate data storage.

Eg: Sensor readings, customer interaction metadata, app log details

Data integration – combining data from different data sources into a unified location

Different organizations employ different forms of data integration but they have a few things in common.

Best practices for data integration are:

Cleanse the data prior to integration

Perform correct ETL mapping to ensure consistency & compatibility

Identification of sources, targets & servers

Advantages of data integration

Makes processes more effective & efficient – because it improves data quality & authenticity, cuts the as various data is integrated people have a one-stop point to access data thereby saving time. Eliminates the need to log into multiple accounts on multiple sites, copying, reformatting, repurposing data before analysis phase.

Improves collaboration among staff – employees from different departments use the same std. data from a unified source rather than different sources.

Provides self-service data accessibility to organizations

Reduces errors & necessity for rework – a unified repository prevents the access of incorrect/outdated data by employees. Most importantly it eliminates the need to redo reporting by synchronizing data & providing means for real-time reporting.

Helps deliver more valuable data to an organization

Data visualization & reporting

Importance of data visualization:

Allows identification of trends & patterns

Helps better understanding of large datasets

Makes it easy to conduct analysis & make predictions

Helps communicate a message or story

Data storytelling – creating a visual, narrative & story behind the data

Data engineering languages

SQL, Java, XML, Python, R