Data Architecture Notes

Contents

1 Introduction **2**

2 Data Type **2**
2.1 Structured data . . . 2
2.2 Unstructured data . . . 3
2.3 Semi-structured data . . . 3
2.4 Short explaination of JSON and XML structures . . . 4
2.5 Semi-Structured data in machine learninig . . . 4

3 Data Warehouse **4**
3.1 Introduction to Data Warehousing . . . 4
3.2 Datawarehousing for data scientists . . . 5
3.3 Cloud datawarehousing . . . 5

4 Data Lake **7**
4.1 Introduction to a data lake . . . 7
4.2 The technology used to build a data lake . . . 7
4.3 Cloud Storage terminology - Buckets and blobs . . . 7

5 Data Lakehouse **7**
5.1 Challenges with the data lake . . . 7
5.2 Intoduction to the data lakehouse . . . 7

6 Data Governance with the Data Mesh **7**
6.1 Intoduction to Data Mesh . . . 7
6.2 Data mesh principles: Domain ownership and data as a product . . . 7
6.3 Data mesh principles: Self service and federated governance . . . 7
6.4 Data Catalog . . . 7
6.5 Data Fabric . . . 7

7 Streaming Data in Data Science **7**
7.1 Introduction to streaming data . . . 7
7.2 Kafka 101 . . . 7
7.3 Lambda architecture . . . 7
7.4 Kapppa architecture and comparison . . . 7
7.5 Word of caution and Resources . . . 7

8 Data infrastrcture for Machine Learning **7**
8.1 Feature Store . . . 7
8.2 Vector Database . . . 7

9 Flowchart and Use case examples **7**
9.1 Data Architecture decision making flowchart . . . 7
9.2 Use case examples and applying the decision . . . 7

1 Introduction

Motivations - Create models that have the potential to deliver high impact on an organisations performance - Many of these models fail to make an impact because you are unable to connect them to the right sources. - Connecting to the right data source is important so that the models train on the right data sets that represent the business problem and thereby improving their accuracy and also removing any bias they might display in production. - Data engineers and data architects are actually responsible for ensuring models get access to the right data. - The reality though is that they require advice and guidance from the data scientists to ensure this does become a reality - Knowing data architecture will also help plan projects in such a way.

Important to plan ahead. Allows for notebook dreams to become highly impactful applications.

Note: Data architecture is outside the core competency of a data scientists roles and responsibility.

2 Data Type

2.1 Structured data

- Structured data typically constitutes only 20% of the total data volume and organisation produces or has access to. – Many organisations have reached a reasonable level of maturity in dealing with structured data. – Its important to understand how structured data is managed by your data engineering team.

Structured data can be found in database tables, CSV files and Excel spreadysssss. - Organised into rows and columns. - The columns contain the features you need to train your model and the rows contain the observations. — Ultimate goal of a data architect or a data engineering team is to extract structured data from the different sources and load it into the data warehouse. —– Since a data warehouse is supposed to have a very accurate information it is not permitted to load data that is unclean or incomplete. —– Hence structured data is first loaded into staging tables where data engineers perform cleansing operation, such as replacing null values in all the columns. —– Staging tables may also be used to join data from different sources and unify them into a single table. —– They may be used to also enrich existing data with information from an external data source.

All in all the staging tables serve as a playground for the data engineer to transform the data and make it ready for loading the data into the warehouse. The data engineer may choose to retain teh staging tables permanently or just generate them dynamically as needed.

Generally speaking, the data scientists will not have access to the staging tables. Once the data is clean unified and enriched, it is read to be loaded into a data warehouse.

Very important: A data warehouse represents the single source of truth for the entire organisation.

Example: Management of a company wants to know total sales for a month. The expect that business analysts can make a query to the data warehouse and get the answer. The data scientist generally has just read access to the data warehouse and is allowed just conditional access to certain tables for exploring the possibility of finding features for data science use cases. Once the features are found the data engineer and not the data scientist transfers the required data into what is called a data mart. The data mart contains a copy of the data that is available in the data warehouse and is refreshed either on request or as per a predefined schedule.

Now, the data scientists usually has full access to the data mart and allowed to do further pre-precessing and manipulation of the contents. Once the data scientist trains the models.

Data marts are again used to store the data that is used for making predictions and the predictions themselves are also stored inside the data mart from where it's picked up by applications or just the dashboarding tools.

2.2 Unstructured data

Most organisations are still learning to extract and derive business value form this type of data. The most common types of unstructured data are images, video, audio and text. Organisations are starting to find a lot of value in analysing text. - There is also computer vision - There is text to speech or speech to text. — Can transcribe conversations into text for further analysis. — Natural language processing is trendy.....ChatGPT etc....

Unstructured data is more than 80% of the data your organisation produces or has access to it needs a storage medium which iis economical easy to write to. This is where the concept of a data lake comes into prominence. Unstructured data typically lands into a data lake. Once data lands in a data lake, it's available for pre-processing and processing by big data frameworks.

Unstructured data needs a lot of processing prior to being useful for data science projects. Since the volume of this data is huge specialised processing frameworks have evolved over time to manage this challenge. – Typical processing activity includes the scaling of images to allow machine learning algorithms to viably learn from them or extraction if text from audio. — For video footage, video annotation is necessary prior to training models to learn from others. — For NLP use cases then tokenization of text is necessary to produce useful models.

To perform all these processing tasks, big data frameworks such as Hadoop, Spark and Apache Bean are commonly used. All these frameworks are capable of reading data from the data lake, processing it and then writing the results back to the data lake. It is important to know that they are capable or reading and writing to the data lake.

NoSQL NoSQL databases are designed to handle a variety of data models and not just tables and rows with columns. They can be used to store unstructured data. — This can become very expensive since data has to be stored on expensive disks for performance reasons. Best Practice: is to keep unstructured data in a data lake and store the metadata in a NoSQL database.

This makes it faster to analyse the properties of the images and fetch only those images that fit the criteria.

2.3 Semi-structured data

This data has some structure but is not confined to just rows and columns. You can define the hierarchy and structure using tags and markers.

— HTML is also a form of semi-structured data where various elements are separated by tags such as title and body (hehehehehe) — Weblogs are generally produced in the JSON format where JSON stands for JavaScript object notation.

Data from sensors is also generally produced in JSON and sometimes in the XML format. Both these formats have implemented tags or markers which help define a structure and this makes it possible to read them.

No two datasets will have the exact same structure or even the same hierarchy.

This makes it easy to write such datasets as developers do not need to confine themselves to a specific structure.

Becomes challenging for those who read such files most of the time, since the strucutre needs to be inferred from the files by the reader. Key Note: Since storing and processing semi structured data from plain text files makes it compute intensive to processe, specialised file formats such as ORC, Parquet and Avro have eveolved. – These file formates are designed

to store data in a columnar format which makes it easier and faster to perform analytics on. – Example, if you with to calculate the total temperatirve from data generated by a sensor it would be easier if the entire temperature was stored in the same file.

Important: If you are dealing with semi structured data in large volumes its best to store them in on of the big data formats.

Semi-structured data can be store in the data lake or in a NoSQL database. – Both are suitable for storing semi-structured data.

Typical flow: - First store the semi-structured data in the data lake and then process it using big data frameworks and then store the data in a NoSQL database. - Once the big data frameworks have cleaned and formatted the data from the data lake, it could then be stored in the NoSQL database. - Storing semi-structured data in a NoSQL database makes it easily accessible by applications and users a like. - Hence, the NoSQL database serves as a good serving layer for the semi strcutured data.

2.4 Short explaination of JSON and XML structures

JSON - Key value pairs. XML - Like HTML but not pretty.

2.5 Semi-Structured data in machine learninig

No machine learning framewokrsk accpet data in a semi strcuted format (see ChatGPT) Semi-structured data is usually flattened into a structured format prior to being used in model training. Flattening is the process of inferring the schema of the dataset and then presenting the required fields in a strucutured format that is in rows and columns. Once the data set is flattened you can use the dataset for training models. — It is not essentially a falttened dataset.

Inference - Do predictions on a deployed model. Most of the model serving frameworks accept JSON as an input and also generally ouput the predictions in the JSON format. Hence no flattening is needed on the inference side of the machine learning workdlow.

3 Data Warehouse

3.1 Introduction to Data Warehousing

A data warehouse is the single source of truth for the entire organisation. A data warehouse is generally used to store information from structured data sources. An organisation collects data from its operational databases and organises it inside a data warehouse, in a format you can perform analytics on. An operational database is one that holds transactions or data that is written by applications. Examples to consider. - Could have a web shop which collects its transactions in a sales database. - CRM which stores all information in a customer database.....

Could perform analytics directly on these operational databases. However, this will impact the performance of your applications. It will be detrimental to your business if you web shop is running slow, just because some data scientist was querying the data to explore the available features. It's also the case that operational database are designed to store transactions and are not easy to query for analytical purposes. (Pretty key point) Operational databases are designed to be written to, whereas analytical databases need to provide more read performance. Analytical databases perform better when the data is stored in a columnar format, whereas operational databases are designed for storing transactions in rows.

Very key point. Makes a lot of sense for any organisation to extract data from all its operational databases and store it in a data warehouse, and store it in a way that makes it easy for querying by analysts and data scientists alike.

Companies also enrich their data warehouse with data from external sources. Combine real time data from the internet with operational data.

3.2 Datawarehousing for data scientists

A data warehouse should have an almost perfect view of the entire organisations structured data estate. Good place to find data to train you machine learning models. Note: If you are struggling to improve the accurary of your model due to lack of feature, it is working looking into the data warehouse to find additional data. Note: Genearlly speaking users only get read access to the data warehouse. Also common to impose restrictions on the visibility of data....don't expect all access from day one.

Most organisations implement a data catalog to provide you with a preview of the data products available in your organisation. A modern enterprise will note let the data scientists query the data directly but instead ask them to query the data directly, but instead ask them to explore the data sets via the data catalog. – Reality is that most enterprises are not there yet so data warehouse skills could be in demand.....

Once the data scientist finds the data needed for training models, he or she will request data engineering team to transfer it to a data mart or even to a feature store. This could be a one time transfer ot a transfer set on a schedule. Example: a data scientist might request the data engineer to schedule the transfer every week so that fresh data is available on a weekly basis for model training. – The data scientists then uses a data science platform or his/her own development environment to engineer new features or pre-process the data for the purposes of model training.

From a data architecture perspective, its important to note that all these tasks are carried out inside the data mart or a feature store as opposed to inside a data warehouse. Once a model is trained, the inference or predictions are also carried out against the data mart or a feature store and not the data warehouse directly.

Data is extracted on a set schedule, which is then sent to the model for predictions. The predictions are also then stored in a data mart which then serves dashboards or applications that utilise the predictive capabilities. Over of what data warehousing typically means and also how data scientists would typically interact with it.

3.3 Cloud datawarehousing

Old way. Data centres. Hardware bundled with proprietary software. Data centres were previously filled with these expensive stacks. On premise data warehouses did perform very well, and they did help organisations derive massive value. Companies need to make heavy upfront investments to buy, install configure and to maintain these systems and to maintain these systems. – There is also no flexibility either. – Had to size the data warehouse for your heavy month end or quarterly reporting jobs. – This meant that hardware was unused for the rest of the time. Can trace back the origin of cloud data warehousing. Cloud data warehouses are hosted natively on public cloud infrastructure as opposed to hardware hosted on premise. – By this, compute memory and storage managed by the cloud providers. – Moreover, they are usually fully managed and hence you don't need staff to manage this infrastructure or perform software updates for that matter. All this is managed for you as a bundled service. Means you can focus on the more important takes, such as loading the data warehouse and extracting the data out of it. What differentiates cloud data warehousing from traditional data warehousing is, number one, the decoupling of compute and storage, which makes it flexible to have different sizes of data warehouses for different needs and that too is on demand. – Is you have to run a monthly reporting on your data, you can spin up a large data warehouse. – Once that report is

done, you can terminate the instance or scale it down to a smaller instance for your day to day needs.

Snowflake is a popular cloud data warehousing technology. Uses t-shirt sizing for its sizing in order to make it easier for users to chose the appropriate size for the appropriate time. Modern cloud data warehousing also support semi-structured data natively. Software vendors are trying to support more and more data types making it easy for customers to have just one technology for all data types. Technology like snowflake, also strives to cater to multiple use cases. Instead of using a separate technology for configuring a data mart, feature store and a data warehouse. The flexibility snowflake provides allows you to use it for all three purposes.

While specialised feature stores do have an edge, snowflake offers to integrate with them and thus offers the best of both worlds.

4 Data Lake

4.1 Introduction to a data lake

4.2 The technology used to build a data lake

4.3 Cloud Storage terminology - Buckets and blobs

5 Data Lakehouse

5.1 Challenges with the data lake

5.2 Intoduction to the data lakehouse

6 Data Governance with the Data Mesh

6.1 Intoduction to Data Mesh

6.2 Data mesh principles: Domain ownership and data as a product

6.3 Data mesh principles: Self service and federated governance

6.4 Data Catalog

6.5 Data Fabric

7 Streaming Data in Data Science

7.1 Introduction to streaming data

7.2 Kafka 101

7.3 Lambda architecture

7.4 Kapppa architecture and comparison

7.5 Word of caution and Resources

8 Data infrastrcture for Machine Learning

8.1 Feature Store

8.2 Vector Database

9 Flowchart and Use case examples

9.1 Data Architecture decision making flowchart

9.2 Use case examples and applying the decision