Practical 1:

What Is Data Science?

Data science is a field of applied mathematics and statistics that provides useful information based on large amounts of complex data or big data.

Data science, or data-driven science, combines aspects of different fields with the aid of computation to interpret reams of data for decision-making purposes.

KEY TAKEAWAYS

- Data science uses techniques such as machine learning and artificial intelligence to extract meaningful information and to predict future patterns and behaviors.
- Advances in technology, the internet, social media, and the use of technology have all increased access to big data.
- The field of data science is growing as technology advances and big data collection and analysis techniques become more sophisticated.

What Is Data Science Useful for?

Data science can identify patterns, permitting the making of inferences and predictions, from seemingly unstructured or unrelated data. Tech companies that collect user data can use techniques to turn what's collected into sources of useful or profitable information.

Cassandra

Cassandra is a distributed database from Apache that is highly scalable and designed to manage very large amounts of structured data. It provides high availability with no single point of failure.

Apache Cassandra is a highly scalable, high-performance distributed database designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. It is a type of NoSQL database. Let us first understand what a NoSQL database does.

NoSQLDatabase

A NoSQL database (sometimes called as Not Only SQL) is a database that provides a mechanism to store and retrieve data other than the tabular relations used in relational databases. These databases are schema-free, support easy replication, have simple API, eventually consistent, and can handle huge amounts of data.

The primary objective of a NoSQL database is to have

- simplicity of design,
- horizontal scaling, and
- finer control over availability.

NoSql databases use different data structures compared to relational databases. It makes some operations faster in NoSQL. The suitability of a given NoSQL database depends on the problem it must solve.

NoSQL vs. Relational Database

The following table lists the points that differentiate a relational database from a NoSQL database.

Relational Database	NoSql Database
Supports powerful query language.	Supports very simple query language.
It has a fixed schema.	No fixed schema.
Follows ACID (Atomicity, Consistency, Isolation, and Durability).	It is only “eventually consistent”.
Supports transactions.	Does not support transactions.

Practical 2:

Homogeneous Ontology for Recursive Uniform Schema

The Homogeneous Ontology for Recursive Uniform Schema (HORUS) is used as an internal data format structure that enables the framework to reduce the permutations of transformations required by the framework. The use of HORUS methodology results in a hub-and-spoke data transformation approach. External data formats are converted to HORUS format, and then a HORUS format is transformed into any other external format. The basic concept is to take native raw data and then transform it first to a single format. That means that there is only one format for text files, one format for JSON or XML, one format for images and video. Therefore, to achieve any-to-any transformation coverage, the framework's only requirements are a data-format-to-HORUS and HURUSto-data-format converter.

Practical 3:
Utilities and Auditing:

Data Processing Utilities:

Retrieve Utilities:

Utilities for this superstep contain the processing chains for retrieving data out of the raw data lake into a new structured format. Build all your retrieve utilities to transform the external raw data lake format into the Homogeneous Ontology for Recursive Uniform Schema (HORUS) data format that I have been using in my projects. HORUS is my core data format. It is used by my data science framework, to enable the reduction of development work required to achieve a complete solution that handles all data formats

Organize Utilities Utilities for this superstep contain all the processing chains for building the data marts. The organize utilities are mostly used to create data marts against the data science results stored in the data warehouse dimensions and facts.

Report Utilities Utilities for this superstep contain all the processing chains for building virtualization and reporting of the actionable knowledge. The report utilities are mostly used to create data virtualization against the data science results stored in the data marts. .

Maintenance Utilities The data science solutions you are building are a standard data system and, consequently, require maintenance utilities, as with any other system. Data engineers and data scientists must work together to ensure that the ecosystem works at its most efficient level at all times.

Backup and Restore Utilities These perform different types of database backups and restores for the solution. They are standard for any computer system. For the specific utilities, I suggest you have an indepth discussion with your own systems manager or the systems manager of your client. I normally provide a wrapper for the specific utility that I can call in my data science ecosystem, without direct exposure to the custom requirements at each customer.

Checks Data Integrity Utilities These utilities check the allocation and structural integrity of database objects and indexes across the ecosystem, to ensure the accurate processing of the data into knowledge.

History Cleanup Utilities These utilities archive and remove entries in the history tables in the databases.

Practical 4:

he Retrieve superstep is a practical method for importing completely into the processing ecosystem a data lake consisting of various external data sources. The Retrieve superstep is the first contact between your data science and the source systems. I will guide you through a methodology of how to handle this discovery of the data up to the point you have all the data you need to evaluate the system you are working with, by deploying your data science skills. The successful retrieval of the data is a major stepping-stone to ensuring that you are performing good data science. Data lineage delivers the audit trail of the data elements at the lowest granular level, to ensure full data governance. Data governance supports metadata management for system guidelines, processing strategies, policies formulation, and implementation of processing. Data quality and master data management helps to enrich the data lineage with more business values, if you provide complete data source metadata.

The Retrieve superstep supports the edge of the ecosystem, where your data science makes direct contact with the outside data world. I will recommend a current set of data structures that you can use to handle the deluge of data you will need to process to uncover critical business knowledge.

Practical 5:

Assess Superstep Data quality refers to the condition of a set of qualitative or quantitative variables. Data quality is a multidimensional measurement of the acceptability of specific data sets. In business, data quality is measured to determine whether data can be used as a basis for reliable intelligence extraction for supporting organizational decision.

Data profiling involves observing in your data sources all the viewpoints that the information offers. The main goal is to determine if individual viewpoints are accurate and complete. The Assess superstep determines what additional processing to apply to the entries that are noncompliant.

Errors

Accept the Error If it falls within an acceptable standard (i.e., West Street instead of West St.), I can decide to accept it and move on to the next data entry

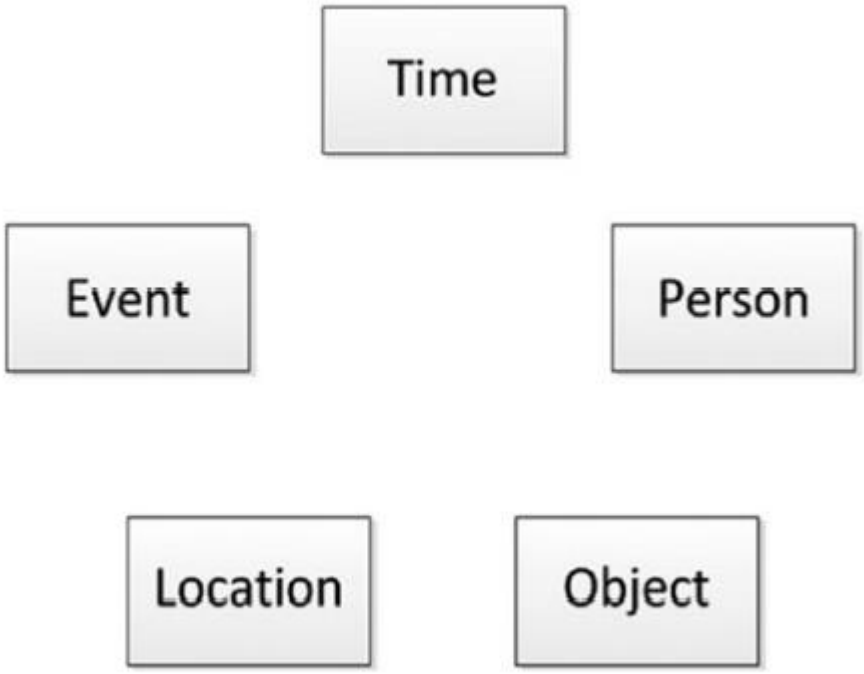
Reject the Error Occasionally, predominantly with first-time data imports, the information is so severely damaged that it is better to simply delete the data entry methodically and not try to correct it. Take note: Removing data is a last resort. I normally add a quality flag and use this flag to avoid this erroneous data being used in data science techniques and algorithms that it will negatively affect.

Correct the Error This is the option that a major part of the assess step is dedicated to. Spelling mistakes in customer names, addresses, and locations are a common source of errors, which are methodically corrected. If there are variations on a name, I recommend that you set one data source as the “master” and keep the data consolidated and correct across all the databases using that master as your primary source.

Create a Default Value This is an option that I commonly see used in my consulting work with companies. Most system developers assume that if the business doesn’t enter the value, they should enter a default value.

Practical 6:

The Process superstep adapts the assess results of the retrieve versions of the data sources into a highly structured data vault that will form the basic data structure for the rest of the data science steps. This data vault involves the formulation of a standard data amalgamation format across a range of projects.



1. *Five categories of data*

sing only these five hubs in your data vault, and with good modeling, you can describe most activities of your customers. This enable you to then fine-tune your data science algorithms, to simply understand the five hubs' purpose and relationships that enable good data science.

Practical 7:

The Transform superstep allows you, as a data scientist, to take data from the data vault and formulate answers to questions raised by your investigations. The transformation step is the data science process that converts results into insights. It takes standard data science techniques and methods to attain insight and knowledge about the data that then can be transformed into actionable decisions, which, through storytelling, you can explain to non-data scientists what you have discovered in the data lake.

The Transform superstep uses the data vault from the process step as its source data. The transformations are tuned to work with the five dimensions of the data vault.

Practical 8:

Organize superstep first, then proceed to the Report superstep. The two sections will enable you, as the data scientist, first to collect the relevant information from your prepared data warehouse, to match the requirement of a specific segment of the customer's decision makers. For example, you will use the same data warehouse to report to the chief financial officer (CFO) and the accountant in Wick, Scotland, but the CFO will receive an overall view, in addition to detailed views of all regions, while the accountant in Wick will see only details related to Wick. This is called organizing your data into a smaller data structure called a "data mart.

Practical 9:

eport Superstep The Report superstep is the step in the ecosystem that enhances the data science findings with the art of storytelling and data visualization. You can perform the best data science, but if you cannot execute a respectable and trustworthy Report step by turning your data science into actionable business insights, you have achieved no advantage for your business.

Practical 10:

What is Power BI?

Power BI is a unified, scalable platform for self-service and enterprise business intelligence (BI). Connect to and visualize any data, and seamlessly infuse the visuals into the apps you use every day.

Easily connect to, model, and visualize your data, creating memorable reports personalized with your KPIs and brand. Get fast, AI-powered answers to your business questions—even when asking with conversational language.