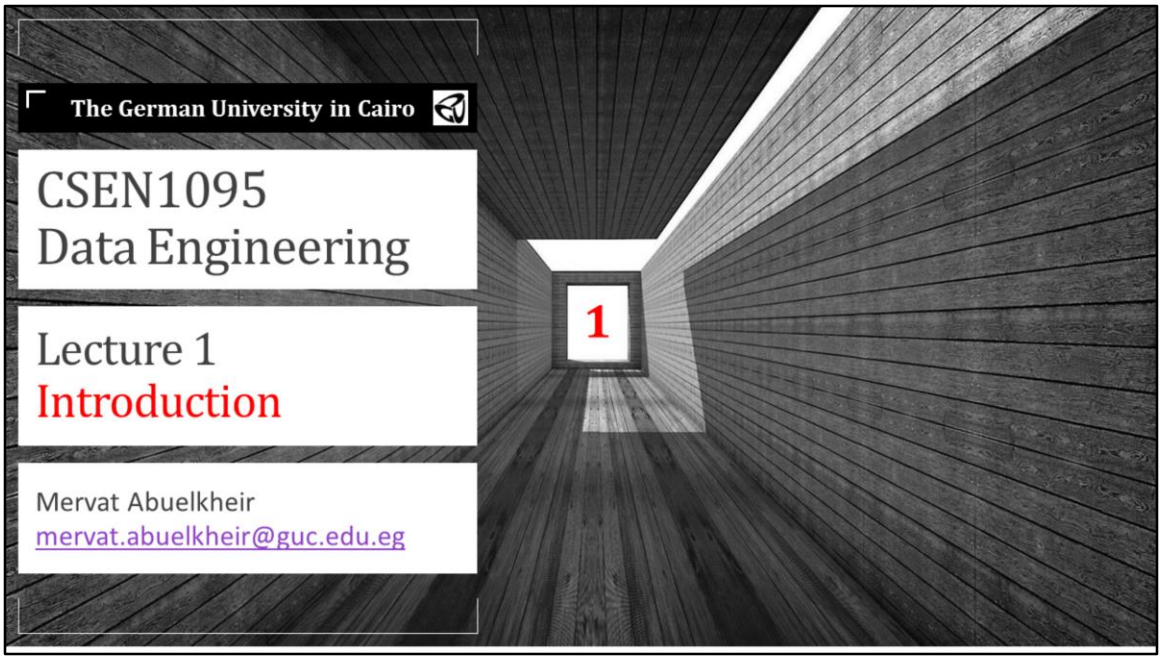
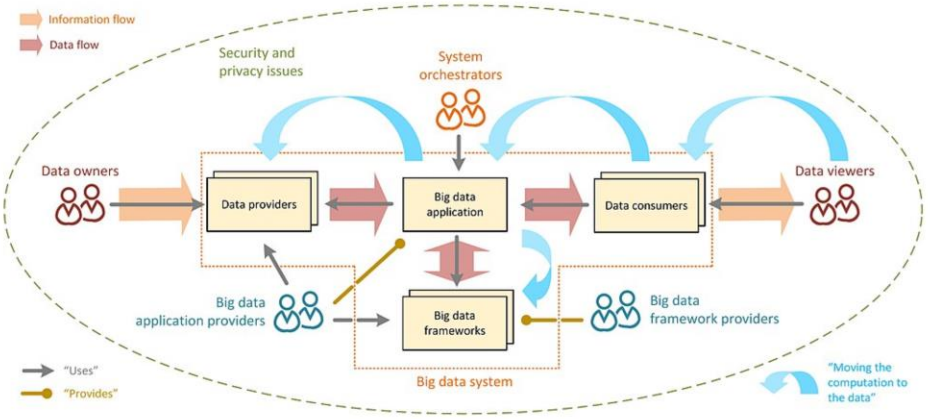

The German University in Cairo
CSEN1095
Data Engineering
Lecture 1
Introduction
1
Mervat Abuelkheir
mervat.abuelkheir@guc.edu.eg

This Course is about Improving Data Quality

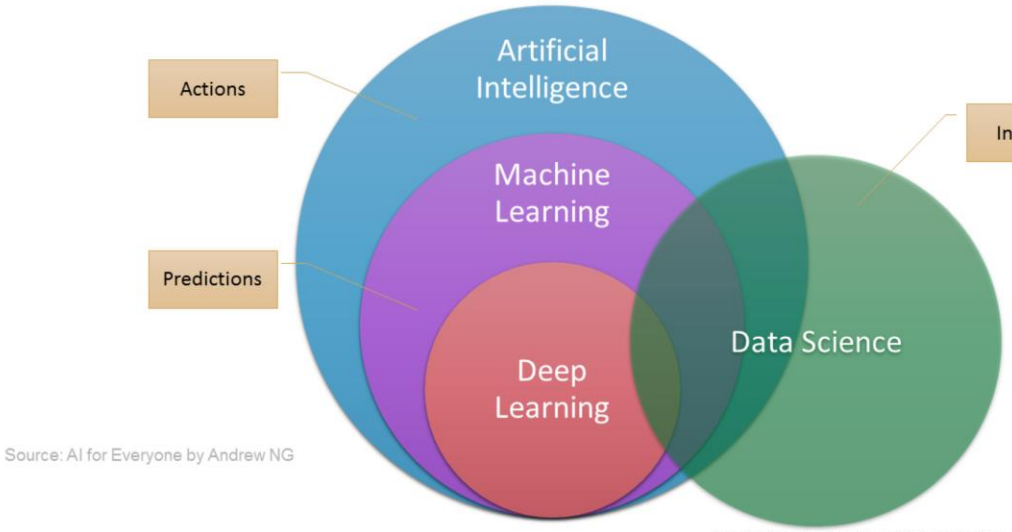
A model is only as good as the data you provide to it

… and Building Big Data Architectures



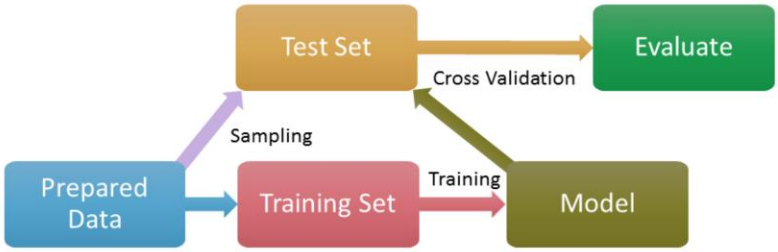
Source: WikiCommons

The Machine Learning Jargon

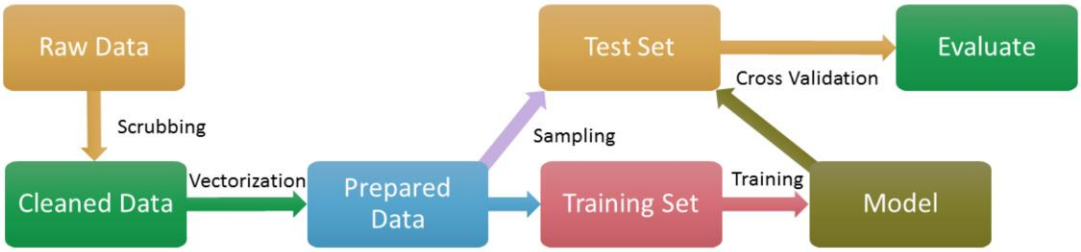


Source: AI for Everyone by Andrew NG

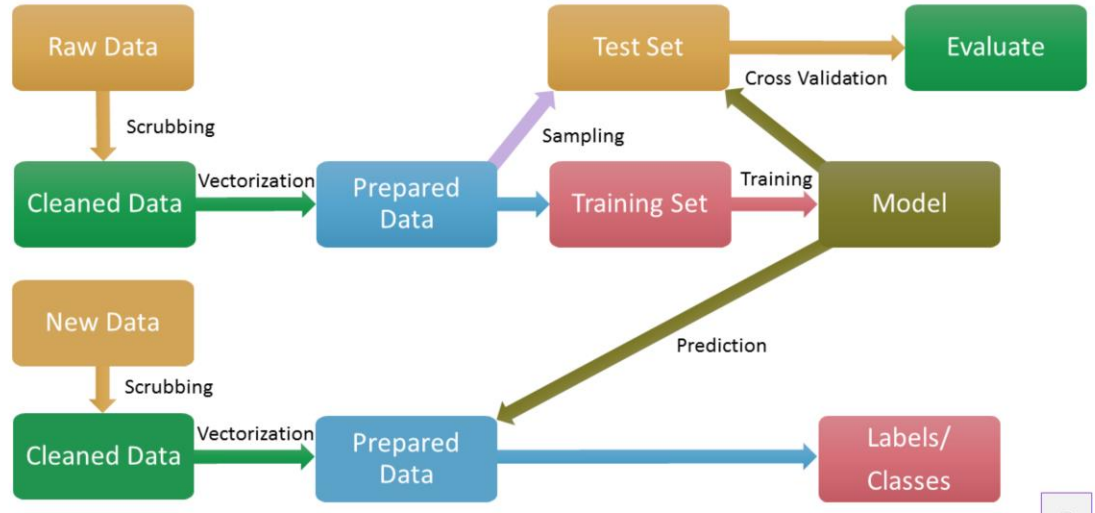
The ML Pipeline – Neat

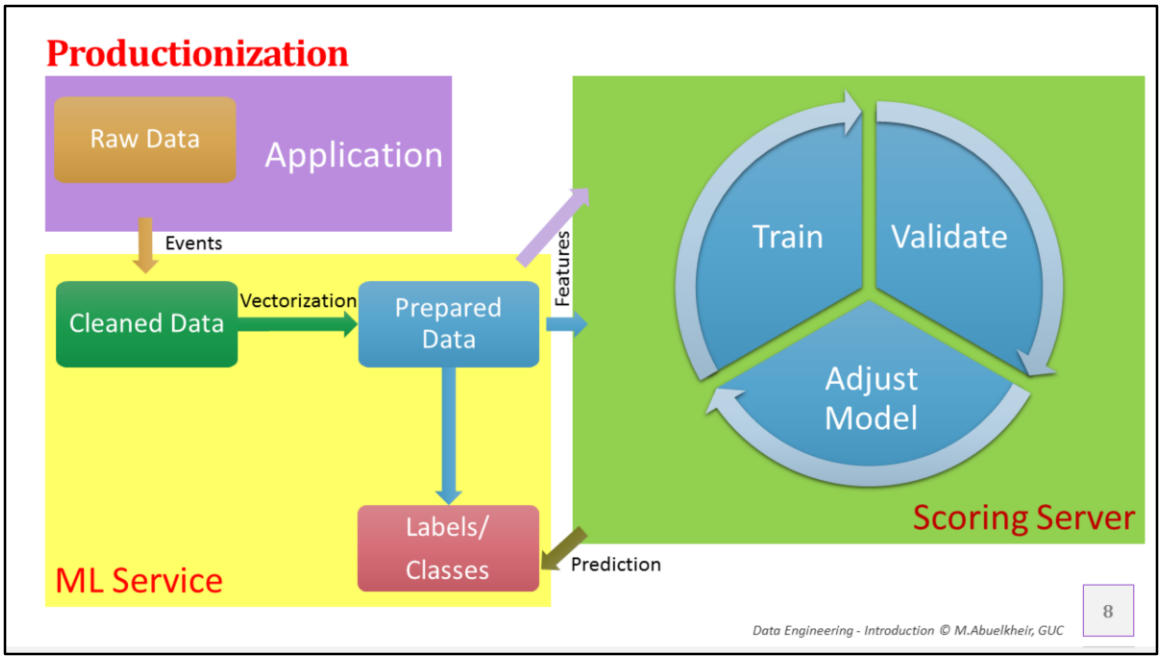


The ML Pipeline – (Not So) Neat

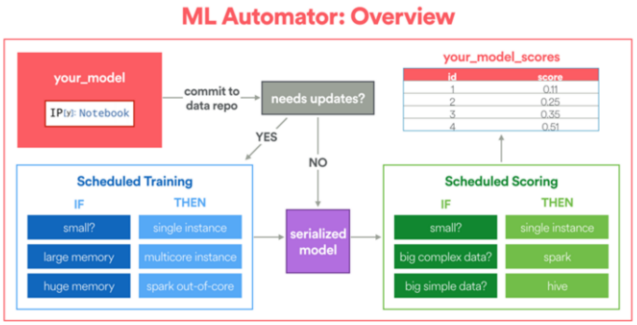


The ML Pipeline – Production

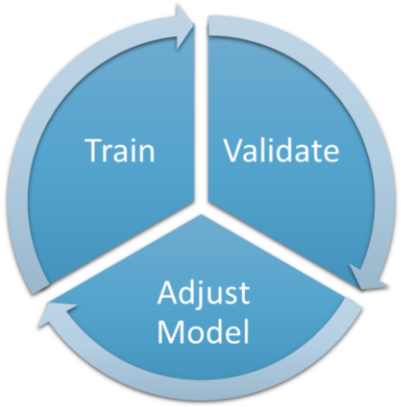



Productionization
Raw Data
Application
Events
Cleaned Data
Vectorization
Prepared Data
Features
Train
Validate
Adjust Model
Scoring Server
Labels/ Classes
Prediction
ML Service
Data Engineering - Introduction © M.Abuelkheir, GUC
8

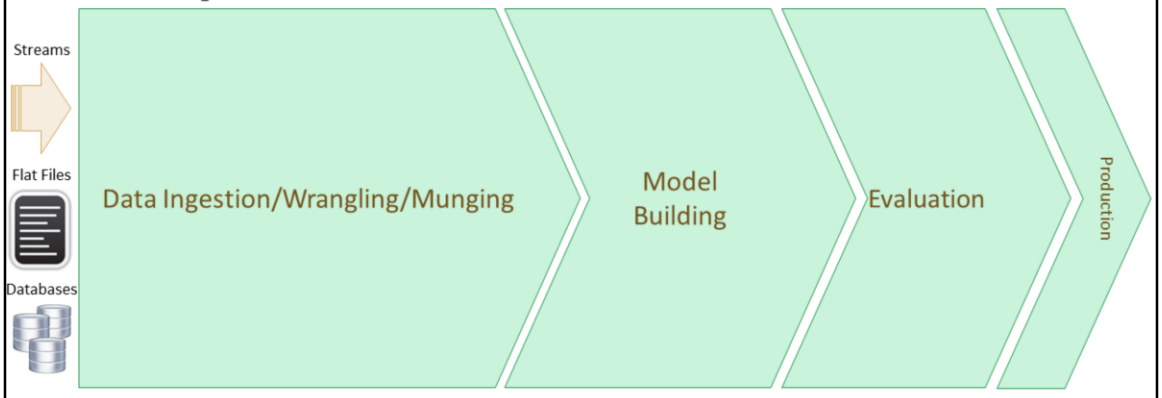
Productionization Example – AirBnB



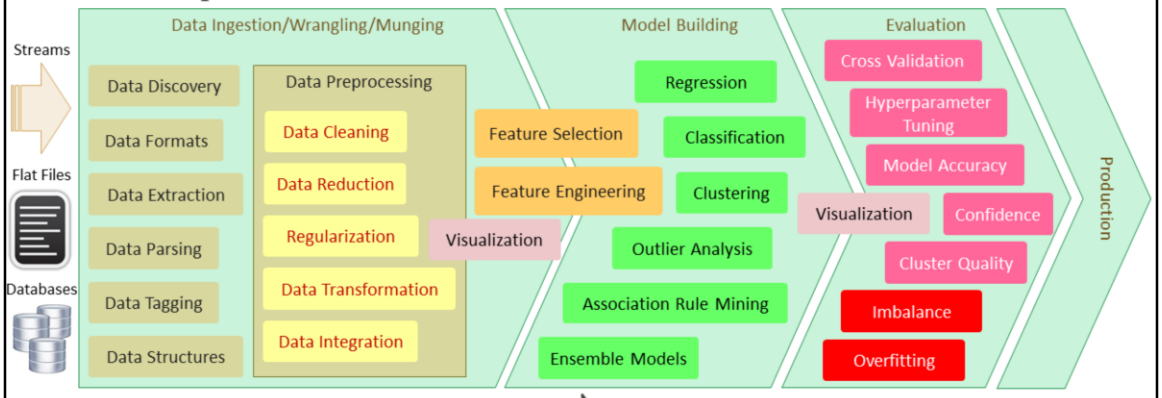
A simplified overview of the *ML Automator* Framework - Airbnb's notebook translation framework



The ML Pipeline – Abstract View

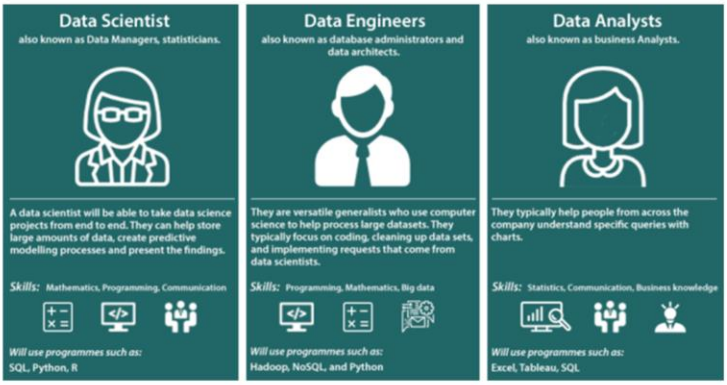


The ML Pipeline – Under The Hood



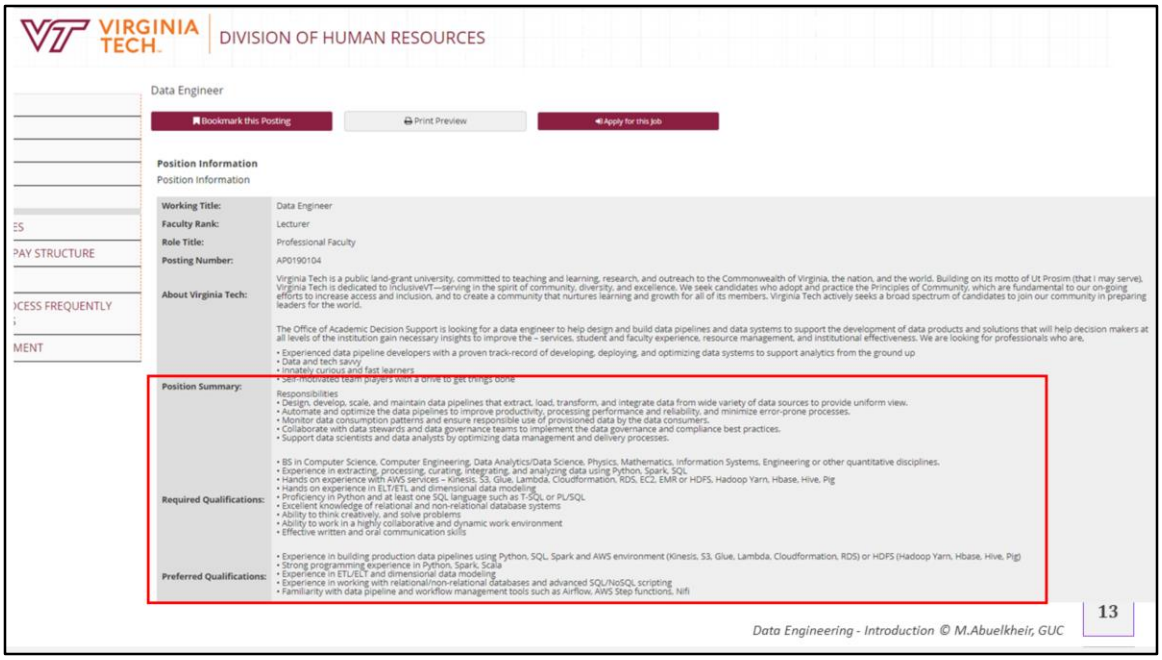
What is a unicorn data scientist?
https://www.quora.com/What-is-a-unicorn-data-scientist

Jobs and whatnot



- Basic Language Requirement: **Python**
- Solid Knowledge of **Operating Systems**
- Heavy, In-Depth Database Knowledge – **SQL** and **NoSQL**
- Data Warehousing – **Hadoop**, **MapReduce**, **HIVE**, **PIG**, Apache **Spark**, **Kafka**
- Basic **Machine Learning** Familiarity

https://www.analyticsvidhya.com/blog/2018/11/data-engineer-comprehensive-list-resources-get-started/



VIRGINIA TECH | DIVISION OF HUMAN RESOURCES

Data Engineer

Bookmark this Posting | Print Preview | Apply for this Job

Position Information

Position Information

Working Title:	Data Engineer
Faculty Rank:	Lecturer
Role Title:	Professional Faculty
Posting Number:	AP0190104
About Virginia Tech:	Virginia Tech is a public land-grant university, committed to teaching and learning, research, and outreach to the Commonwealth of Virginia, the nation, and the world. Building on its motto of Ut Prosim (that I may serve), Virginia Tech is dedicated to inclusiveVT—serving in the spirit of community, diversity, and excellence. We seek candidates who adopt and practice the Principles of Community, which are fundamental to our on-going efforts to increase access and inclusion, and to create a community that nurtures learning and growth for all of its members. Virginia Tech actively seeks a broad spectrum of candidates to join our community in preparing leaders for the world. The Office of Academic Decision Support is looking for a data engineer to help design and build data pipelines and data systems to support the development of data products and solutions that will help decision makers at all levels of the institution gain necessary insights to improve the – services, student and faculty experience, resource management, and institutional effectiveness. We are looking for professionals who are, • Experienced data pipeline developers with a proven track-record of developing, deploying, and optimizing data systems to support analytics from the ground up • Data and tech savvy • Innately curious and fast learners • Self-motivated team players with a drive to get things done
Position Summary:	Responsibilities • Design, develop, scale, and maintain data pipelines that extract, load, transform, and integrate data from wide variety of data sources to provide uniform view. • Automate and optimize the data pipelines to improve productivity, processing performance and reliability, and minimize error-prone processes. • Monitor data consumption patterns and ensure responsible use of provisioned data by the data consumers. • Collaborate with data stewards and data governance teams to implement the data governance and compliance best practices. • Support data scientists and data analysts by optimizing data management and delivery processes.
Required Qualifications:	• BS in Computer Science, Computer Engineering, Data Analytics/Data Science, Physics, Mathematics, Information Systems, Engineering or other quantitative disciplines. • Experience in extracting, processing, curating, integrating, and analyzing data using Python, Spark, SQL • Hands on experience with AWS services – Kinesis, S3, Glue, Lambda, Cloudformation, RDS, EC2, EMR or HDFS, Hadoop Yarn, Hbase, Hive, Pig • Hands on experience in ELT/ETL and dimensional data modeling • Proficiency in Python and at least one SQL language such as T-SQL or PL/SQL • Excellent knowledge of relational and non-relational database systems • Ability to think creatively, and solve problems • Ability to work in a highly collaborative and dynamic work environment • Effective written and oral communication skills
Preferred Qualifications:	• Experience in building production data pipelines using Python, SQL, Spark and AWS environment (Kinesis, S3, Glue, Lambda, Cloudformation, RDS) or HDFS (Hadoop Yarn, Hbase, Hive, Pig) • Strong programming experience in Python, Spark, Scala • Experience in ETL/ELT and dimensional data modeling • Experience in working with relational/non-relational databases and advanced SQL/NoSQL scripting • Familiarity with data pipeline and workflow management tools such as Airflow, AWS Step functions, Nifi

As a Data Engineer, you will:

- Develop processes supporting data transformation, data structures, metadata, dependency and workload management.
- Develop new data integrations as well as maintaining the existing ones.
- Create and maintain the optimal data model across the pipeline components, and track the model consistency across the pipeline.
- Develop REST APIs over data models to facilitate and unify the data access layer.
- Participate in defining the most suitable backing technologies (relational, NoSQL, in-memory, queues) for Agolo's pipeline components and defining the scaling challenges and factors.
- Build analytics tools that utilize the data pipeline to provide actionable insights into customer acquisition, operational efficiency, and other key business performance metrics.
- Assemble large, complex data sets that meet functional / non-functional business requirements.

Required qualifications:

- Passionate about building and optimizing data pipelines, architectures, and data sets.
- At least 2+ years of relevant experience.
- Proven track record in the following:
  - Object-oriented/functional scripting languages: Java, Python, Scala, JavaScript.
  - SQL database.
  - Algorithmic and problem-solving skills.
  - Strong understanding of software engineering concepts and design patterns.
- Clean code/design advocate.
- Professional-level English written and oral communication skills.

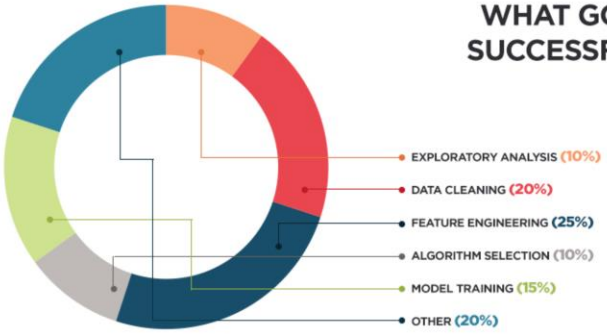
Experience with any of the following will be a great plus:

- Knowledge of Distributed Systems.
- NoSQL databases, any of Elastic, Solr, MongoDB, Redis, CouchDB.
- Knowledge with Big data tools: Hadoop, Spark, Kafka, etc.
- Any cloud infrastructure: Azure, GCP, AWS.
- Stream-processing systems: Kafka-Streaming, Storm, Spark-streaming.
- Analytical skills related to working with unstructured dataset.
- Knowledge with DevOps tools and technologies: CI/CD CircleCI, ELK, Docker containers, Kubernetes, monitoring and alerting tools (e.g. Prometheus and Grafana).

Why you should join the team:

- We offer very competitive salaries in USD.

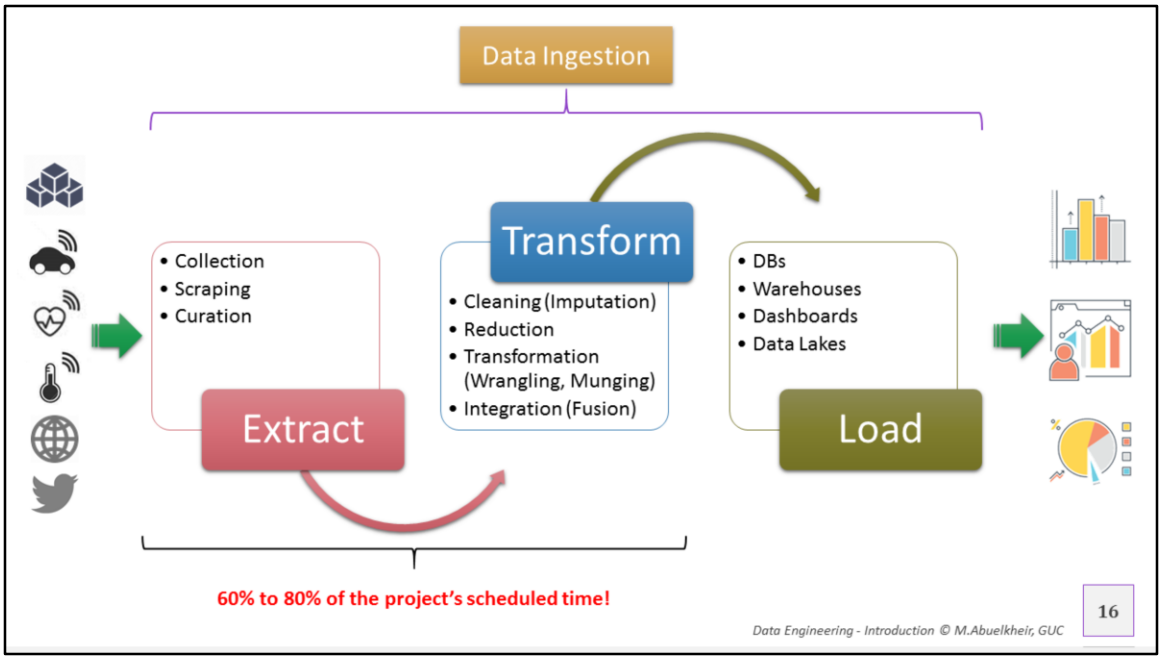
WHAT GOES INTO A SUCCESSFUL MODEL



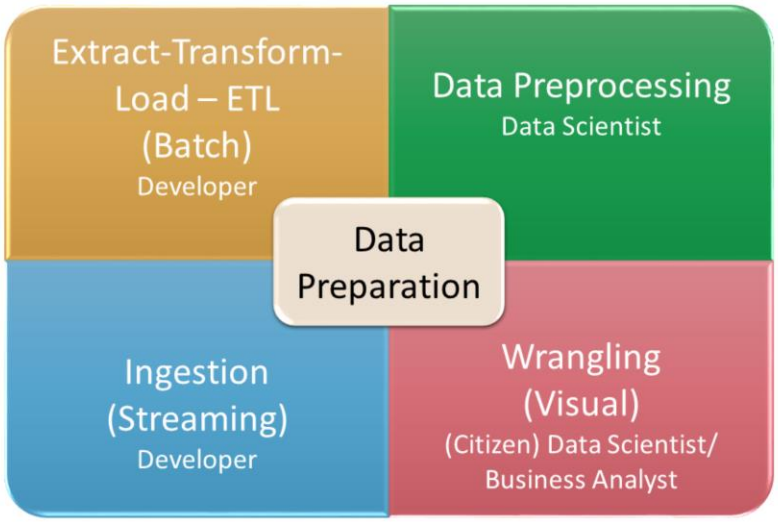
DATA SCIENCE PRIMER
by EliteDataScience

- Data workers spend more than 40% of their time searching for and preparing data instead of gleaning insights
- On average, data workers leverage more than six data sources, 40m rows of data and seven different outputs along their analytic journey

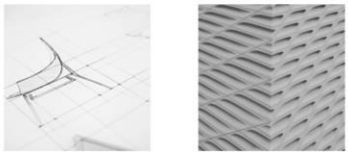
https://www.information-age.com/productivity-in-data-science-123482699/
https://www.information-age.com/data-engineer-sexiest-job-21st-century-123480578/



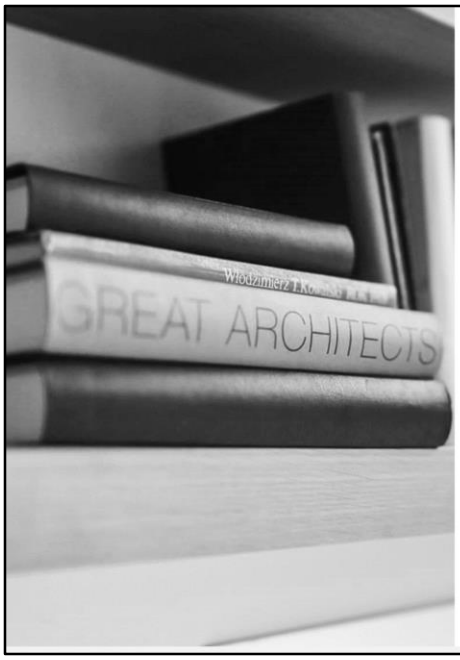
ETL is essentially a blueprint for how the collected raw data is processed and transformed into data ready for analysis


Extract-Transform-Load – ETL (Batch)
Developer
Data Preprocessing
Data Scientist
Data Preparation
Ingestion (Streaming)
Developer
Wrangling (Visual)
(Citizen) Data Scientist/ Business Analyst

Week		Saturday Lecture	Monday Lecture
1	Data Quality	Introduction	
2		Data exploration • Understand data • Use statistics and visualizations	Familiarize yourselves with Python Download Anaconda and Pandas
3		Data cleaning and imputation • Missing values • Noise and Outliers	NumPy – numeric vector and matrix operations Matbloblib and Seaborn - visualization
4		Data transformation • Transformation	Explore pandas and DataFrames Explore pyjanitor Put data in tidy format
5		Data reduction • Sampling/Partitioning • Feature selection • Feature engineering	Normalization Scikit-learn – feature transformation Download and install Spark
6	Data Architecture	Data integration	Sampling, partitioning, regression Training, testing, validation set creation
7		Introduction to the Hadoop ecosystem	Ingestion, scrapping, merge, join, correlation
8		Data storage and management • SQL or NoSQL?	HBase – NoSQL Use ORM – SQLAlchemy to insert data
9		Building the data pipeline I	Hive – the data warehouse for big data
10		Building the data pipeline II	Kafka
11		Building data streaming architectures	Spark and Spark streaming



Course Outline

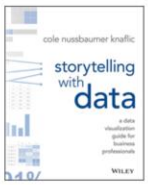


Books

Feature Engineering for Machine Learning

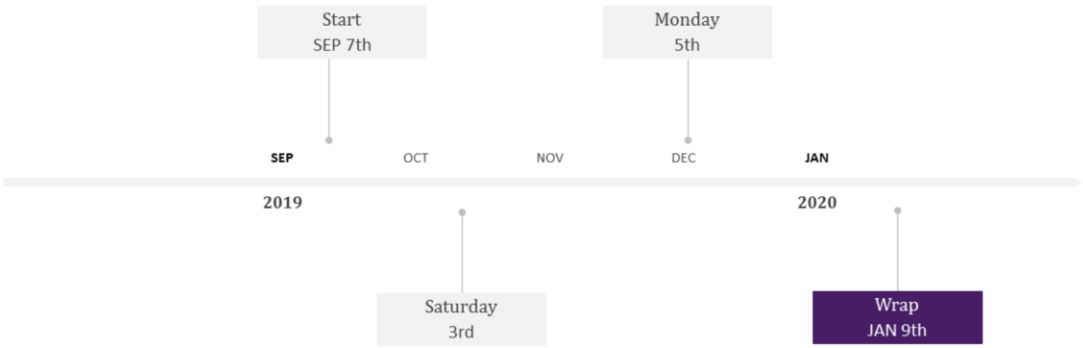
Python Data Science Handbook

Designing Data-Intensive Applications

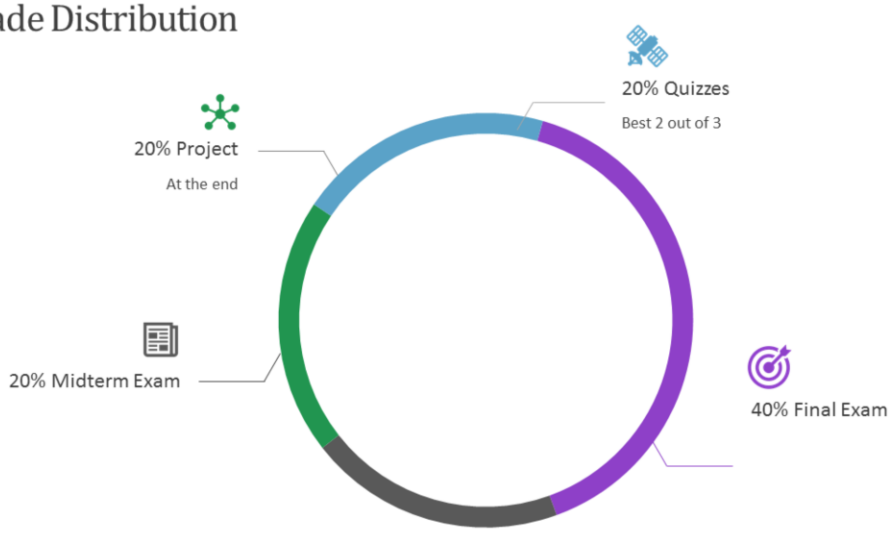


Timeline

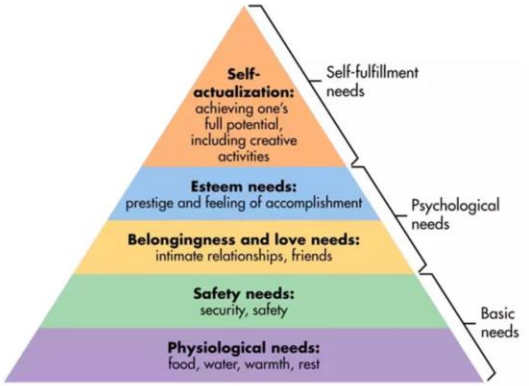
Office Hours: Sunday 13:00 – 16:00 (would prefer to send email in advance)



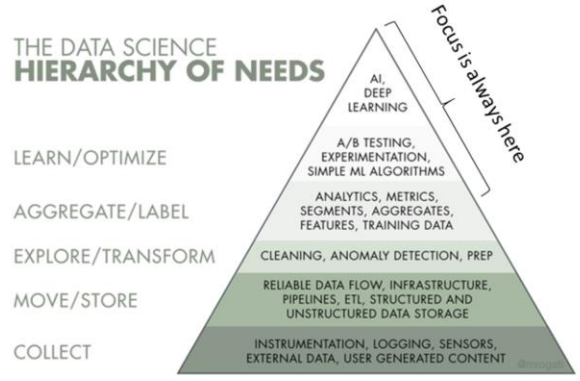
Grade Distribution



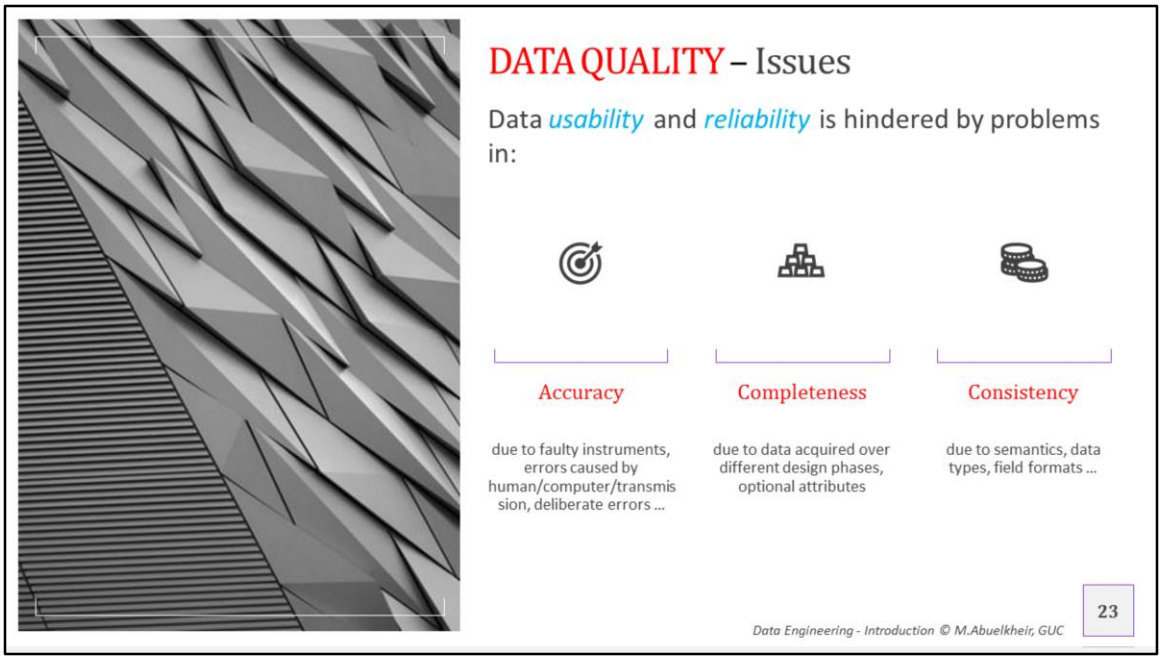
What We Need to Do …



Source: Saul McLeod's SimplyPsychology post "Maslow's Hierarchy of Needs"



Source: Monica Rogati's Medium post "The AI Hierarchy of Needs"



Accuracy: The values contained in each field of the database record should be correct and accurately represent "real world" values.

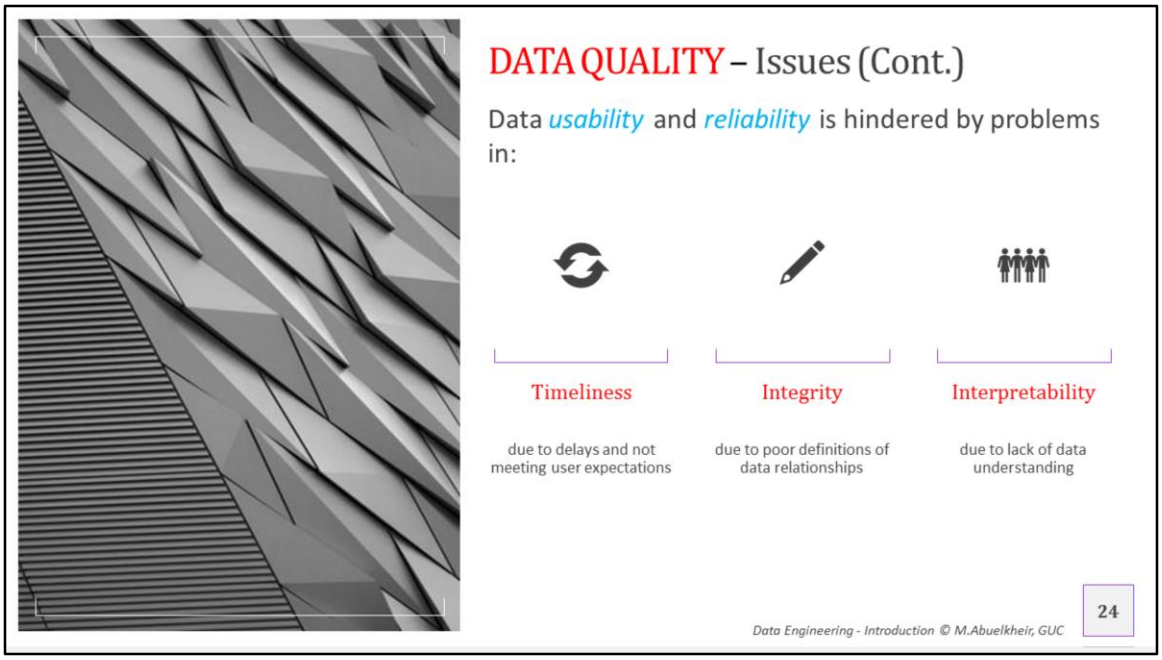
Example: A recorded address should be a real address. Names should be spelled correctly.

Completeness: The data should contain all the necessary and expected information, and the scope of the data element should be understood by the user. No required elements should be missing or in an unusable state.

Example: If first and last name are required in a form, but middle name is optional, the form can still be considered complete if no middle name is entered.

Consistency: Recorded data should be the same throughout the organization and across all systems. Watch out for conflicting information between data sets, records, and systems.

Example: Data for a sale recorded in the company's CRM should match data recorded in the financial software.



Conformity: Data should conform to certain standards of type, size, format, etc.

Example: All dates should be in mm/dd/yyyy format. Names should use only letters, not numbers or symbols.

Timeliness: The data should be available when it's expected and needed by the user. Whether data is timely depends on user expectations.

Integrity: The data should be valid across relationships, meaning that there are recorded relationships that connect all the data together. Note that unlinked records may introduce duplicate entries in your system.

Example: If you have an address recorded in your database, but there is no person, company, or other relationship associated with the address, the data is invalid. It is an orphaned record.

Interpretability: One real-world entity should correspond to only one thing in your data. Duplicate entries should be eliminated.

Example: If you have a company record with the name "Salesforce" and another with the name "SalesForce," one record should be deleted (ideally the one that doesn't reflect Salesforce's preferred capitalization).

Data Preprocessing

Improve the quality of data and make it suitable for the requirements of the intended use

Data cleaning

Fill in missing values, smooth noisy data, identify/remove outliers, resolve inconsistencies

Data transformation

Change how the data looks, adjust the scale of the data

Data integration

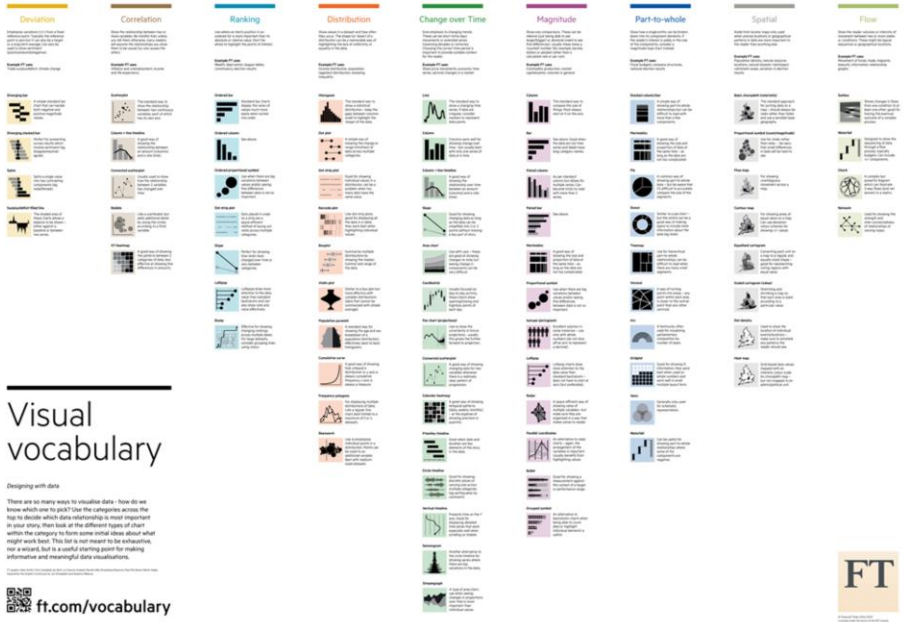
Include data from multiple sources, map semantic concepts, infer attributes

Data reduction

Obtain a reduced representation of the data that is much smaller in volume

Data Visualization

The Golden Tool



Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

ft.com/vocabulary

DATA ARCHITECTURE: Why Do We Need One?

Big data is …

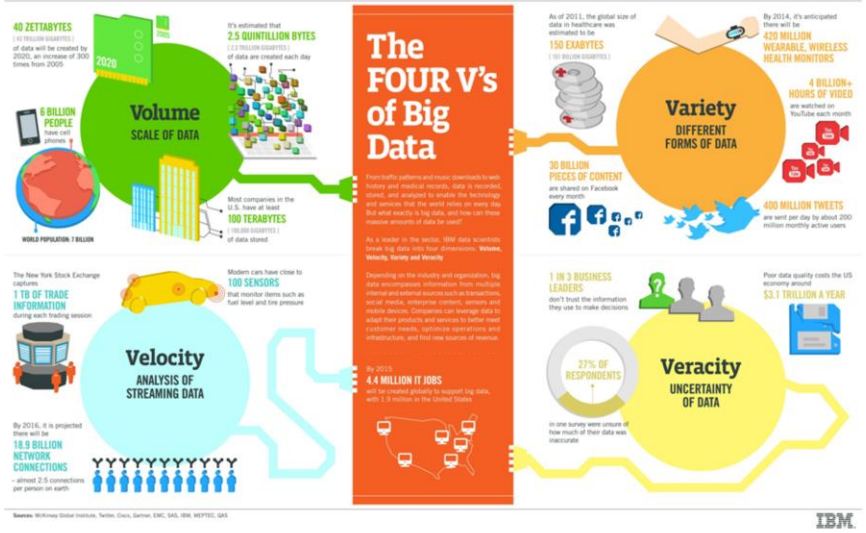
- "…data that is an order of magnitude bigger than you're accustomed to."

  –Doug Laney, Gartner

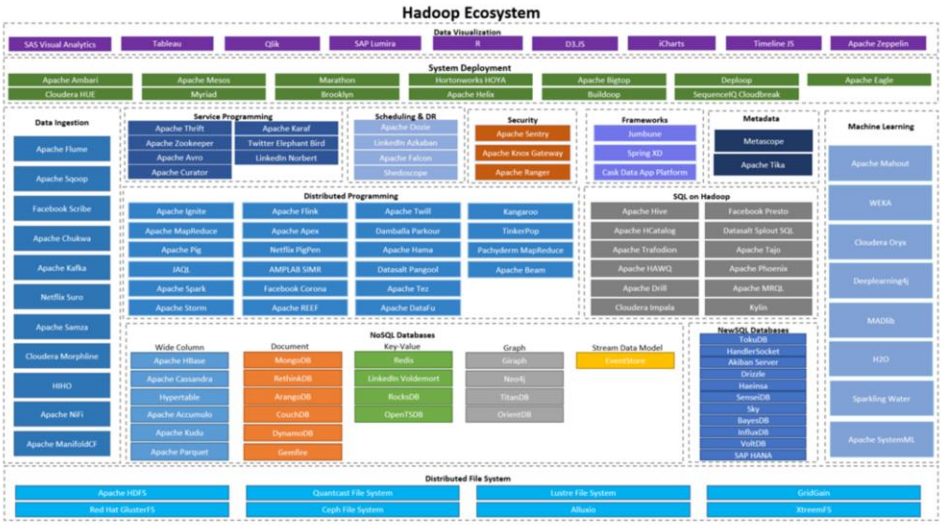
- "… data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures."

  –Ed Dumbill, Program Chair, O'Reilly Strata Conference

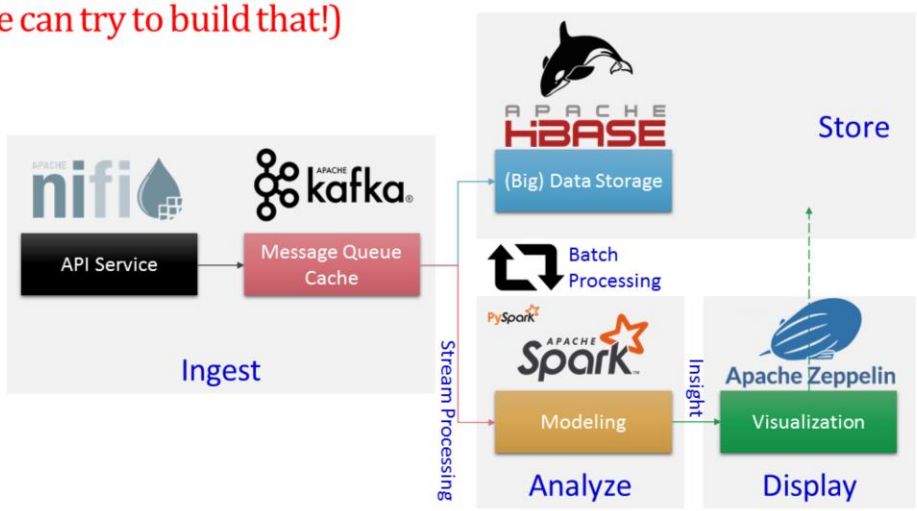
DATA ARCHITECTURE



DATA ARCHITECTURE – The Hadoop Ecosystem

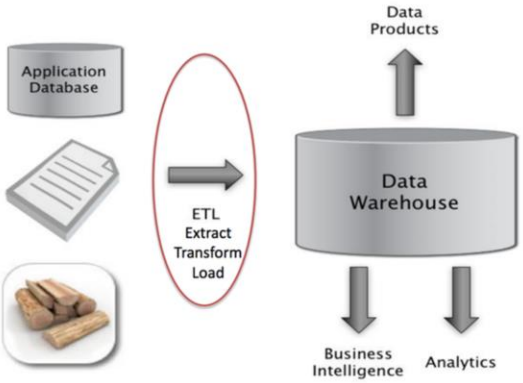


An Example Modern Data Pipeline
(We can try to build that!)



ETL

- **Why do we need a data warehouse** if we already have an application/production database?
  - Different results for same query (production DB reflects system's dynamics)
  - Delays to the real system
  - Outages



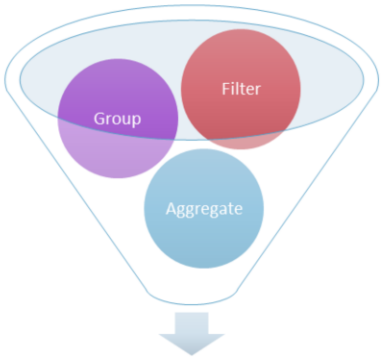
Source: Jeff Hammerbacher's slide from UC Berkeley CS 194 course

E – Data Extraction

- **Obtain data from different sources and import it into database/data warehouse**
- Emphasis is on rapid extraction and delivery
- Data can be streamed (ingested in real time)or batched (ingested in discrete chunks periodically)
- Diverse formats and source volume dictate different ingestion mechanisms (scale)

T – Data Transformation (Preprocessing)

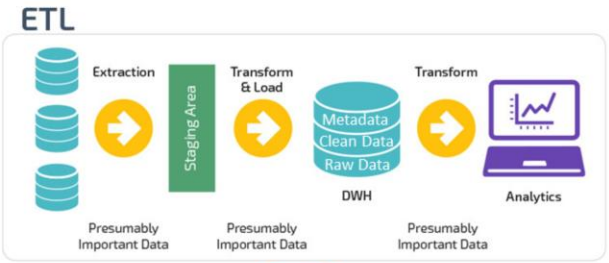
- Half the course is about this phase!
- Data collected is data raw
- **Data needs to be filtered and cleaned**
- **Data from different sources needs to be integrated**
- **Data with large volumes needs to be scaled down or partitioned**

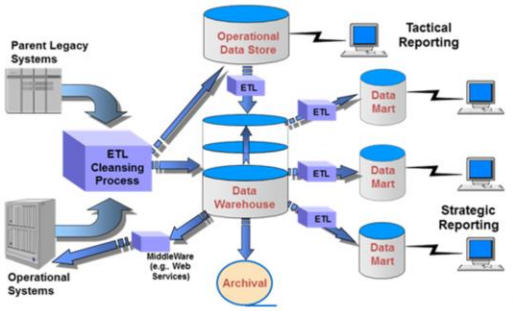


Analysis-ready dataset

L – Data Loading

- **Transport data to final destination**
- Either consumed directly by end users
- Or fed to another ETL job in the flow
  - Data lineage
  - Data provenance
- ETL orchestration engine (e.g. Airflow)





The Problems





Problem	Predict prob. of recovery from a disease	Predict production of crop yield over 5 years	Predict global market demand for crops over 5 years	Characterize performance for football players	Predict SME growth
Variables	Vital stats Past history Genetic propensity ...	Climate Fertilizers available Soil type ...	Government tenders Exports Crop cycles ...	Goals Minutes per match Yellow cards ...	Earnings Market share Valuation ...

how many variables can be measured and how many are a part of the available data? How many are correlated and redundant? Which values are suspicious and possibly inaccurate?

Example research investigating the problems:
Problem 1: https://www.sciencedirect.com/science/article/pii/S2214782917300751
Problem 2: https://www.mdpi.com/2220-9964/8/5/240/pdf
Problem 3: http://www.amis-outlook.org/fileadmin/user_upload/amis/docs/Market_monitor/AMIS_Market_Monitor_current.pdf
Problem 4: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0211058
Problem 5: https://www.emerald.com/insight/content/doi/10.1108/JFBM-09-2017-0029/full/html

GUC

Thank You