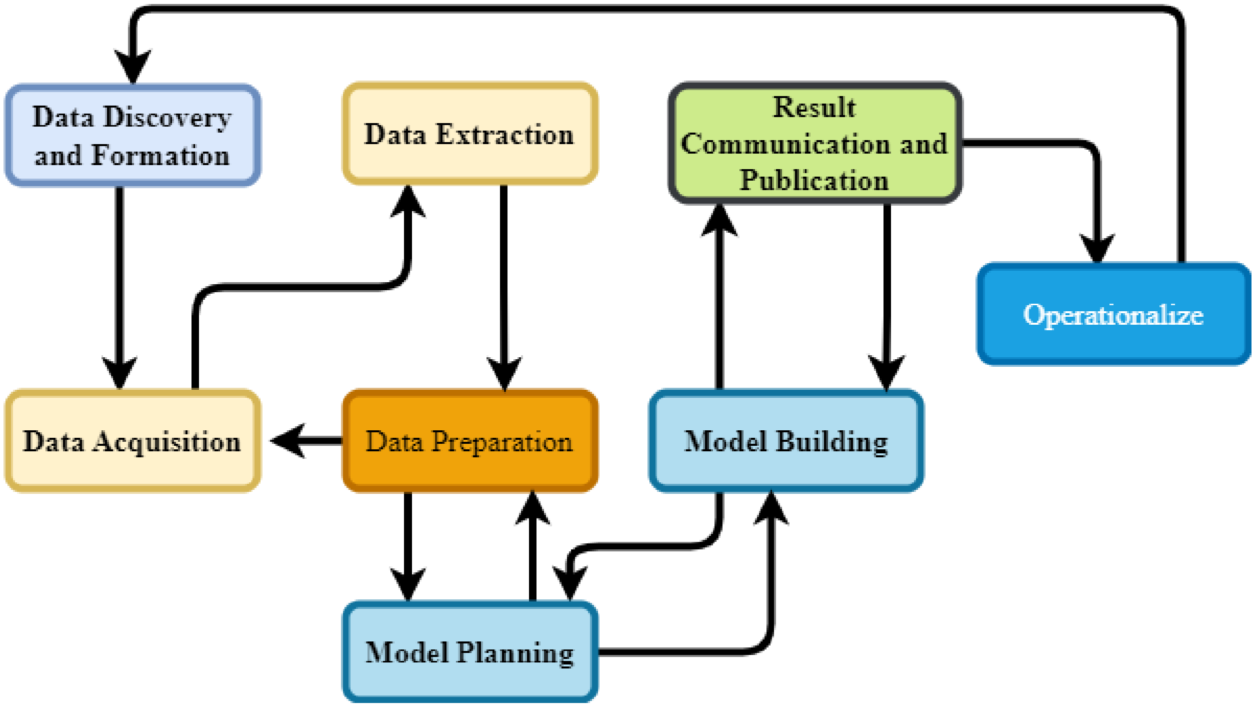
Big Data Life Cycle

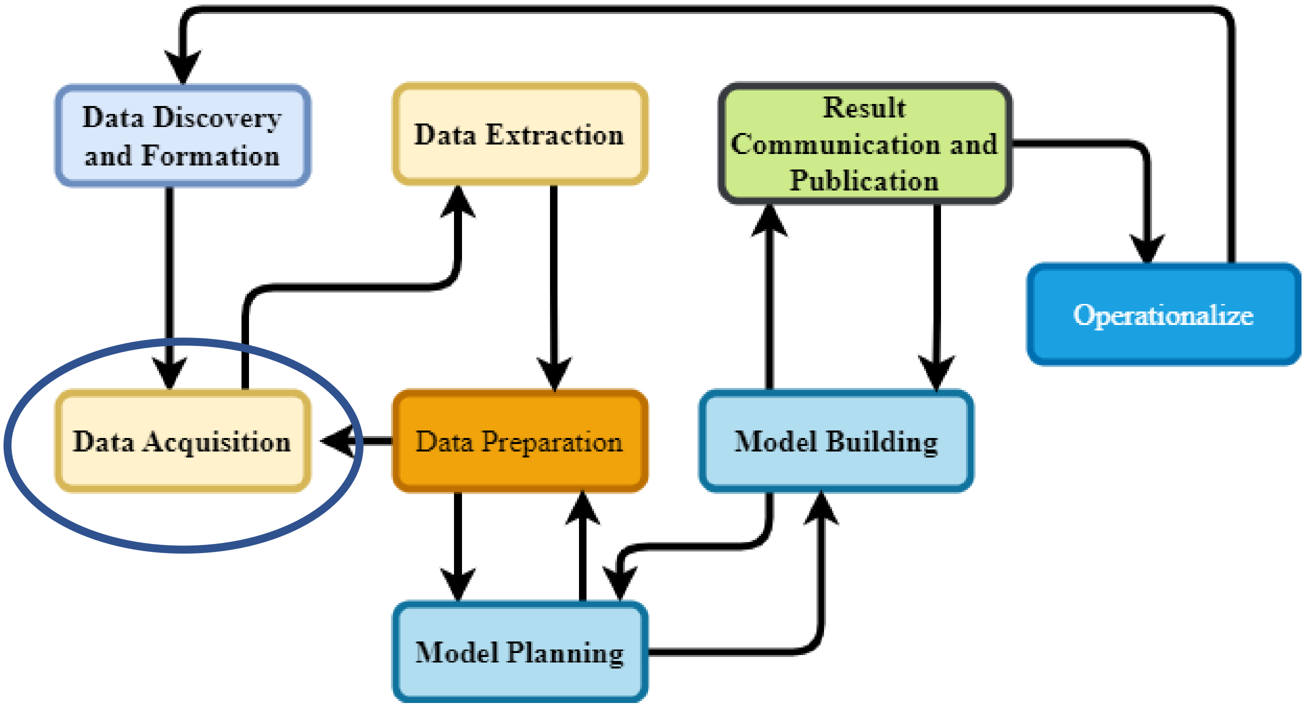
- **Overview:**
  - The Data Analytics Lifecycle outlines how data is created, gathered, processed, used, and analyzed to meet corporate objectives.
  - It provides a structured method of handling data so that it may be transformed into knowledge that can be applied to achieve business growth.
  - Big data lifecycle differs from traditional lifecycle due to velocity and volume.
  - It is a repetitive set of steps that you need to take to complete and deliver a project to your client.
  - Different big data projects require different processing steps.
  - Types of big data analytics objectives:
    - Classification: How can categorize this new customer/patient/etc.?
    - Clustering: Which group of customers will buy this new product?
    - Intrusion Detection: This is not a normal transaction?
    - Recommendation: What should we offer this type of customers? Which option should be taken?
    - Regression: what will sales look like over the next six months?

- **What are big data lifecycle Steps?**

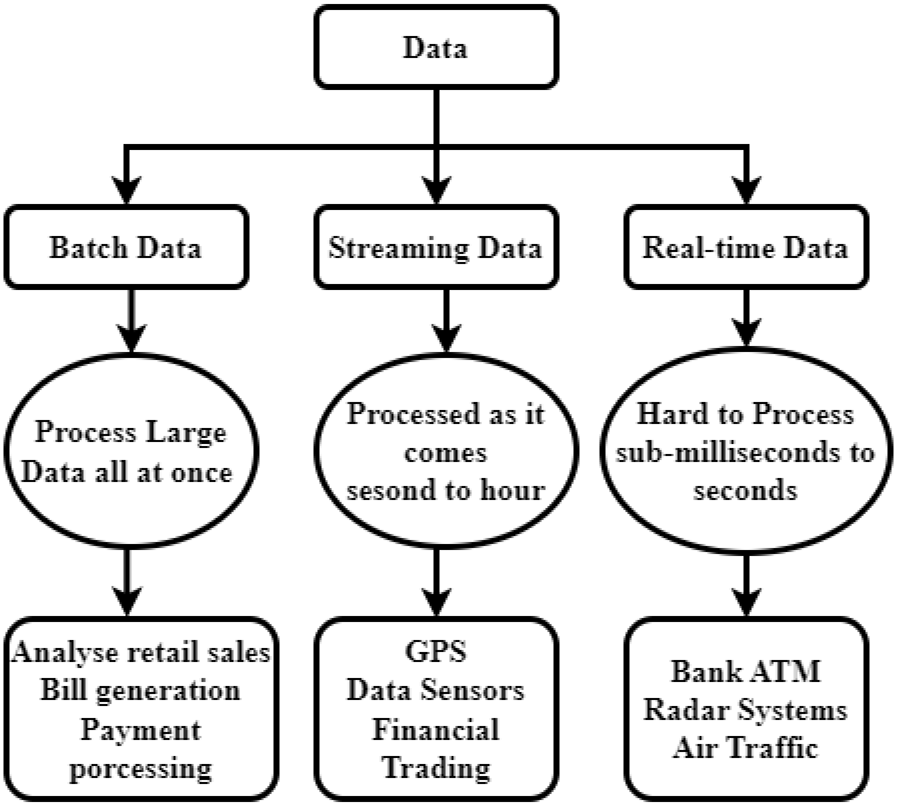


  - **Phase 1: Data Discovery and Formation**
    - Understanding the Business Problem
    - Understanding the objectives of the analysis
    - Speak to stakeholders to understand the business problem that the client is facing.

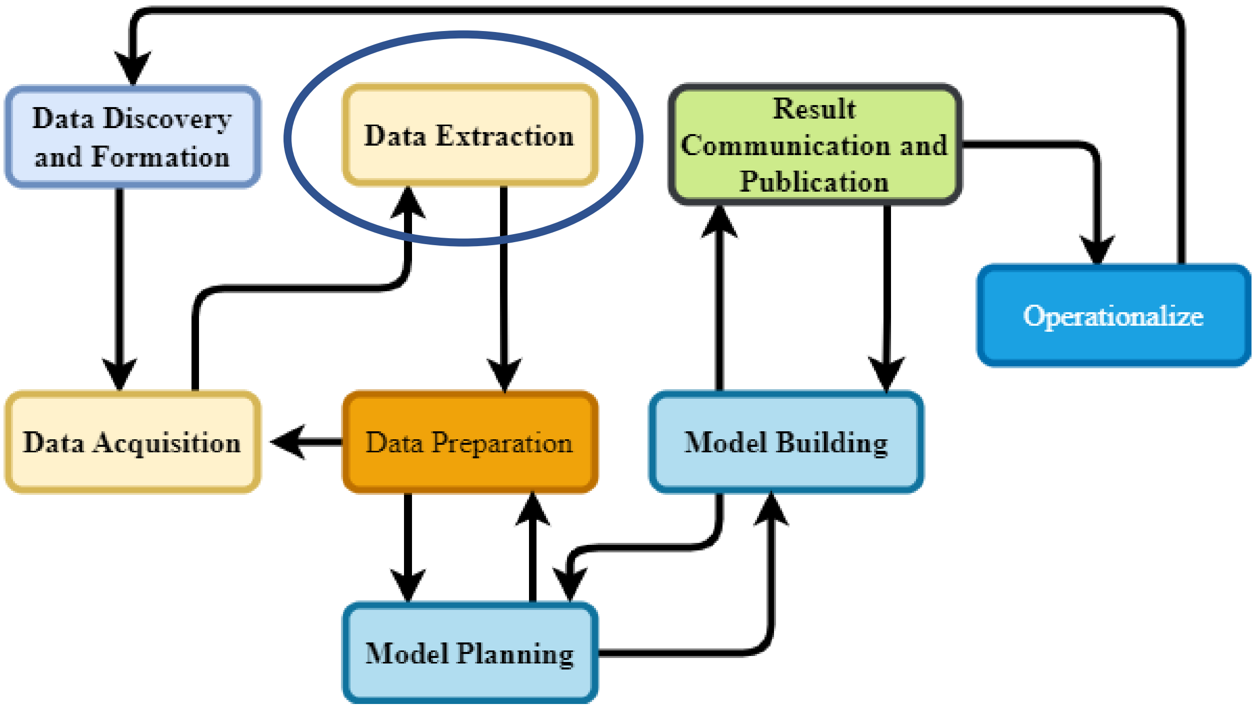
- **Phase 2: Data Acquisition**



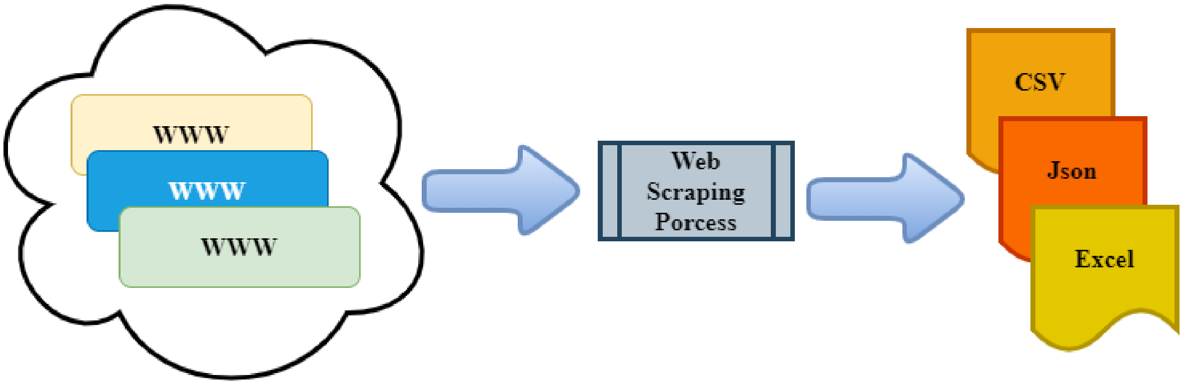
  - Identify various data sources
  - Collect data from both internal and external sources related to the objectives of the analysis.
  - Two types of data:
    - Batch data
    - Run-time/Streaming data
      - Examples: Web server logs, social media posts, APIs data, web scraping, or internal information such as excel spreadsheets, pdf reports, etc.


Data
Batch Data
Streaming Data
Real-time Data
Process Large Data all at once
Processed as it comes sesond to hour
Hard to Process sub-milliseconds to seconds
Analyse retail sales
Bill generation
Payment porcessing
GPS
Data Sensors
Financial Trading
Bank ATM
Radar Systems
Air Traffic

- **Phase 3: Data Extraction**



  - Extract critical data from documents such as PDF and scanned invoices and contract documents.
  - Parsing delimited textual data such as web log. Note some big data tool can directly process this type of files.
  - Requires different solutions for different types of data: batch vs. stream. The data schema may change over time: schema on write vs. schema on read.
  - Web Data Scraping:
    - Monitor the pricing of products
    - Monitor product feedback and reviews



- Examples:
    - Extract data from a Json or XML file

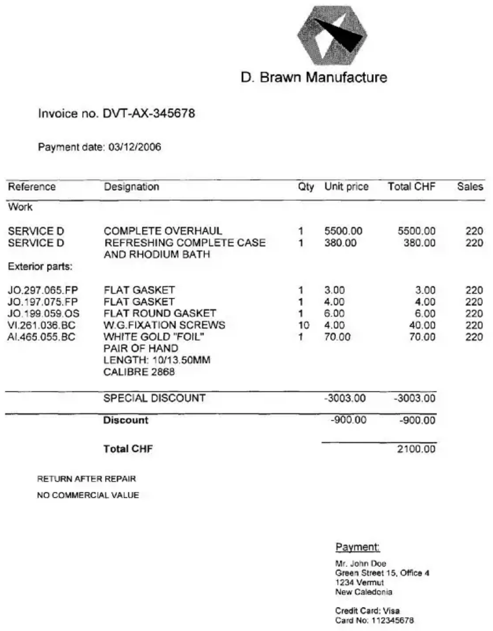
```
{
    "_id" : EmpId("A1101"),
    “FirstName" : "James",
    LastName" : "Berry",
    Age" : 51,
    Interests" : [ "Spinning",  “Soccer” ]
    “Address": {
            Street": "000 Leesburg Pike",
        "City": "Falls Church",
        "State": "VA “
        "Zip": "22041”,
    }
}
```

ID	FirstName	LastName	Age	InterestID	AddressID

OR

ID	FirstName	LastName	Age	Interest	Address_St	Addr_City	Addr_State	Addr_zip

- Example: Invoice/PDF files (towardsdatascience.com)



D. Brawn Manufacture

Invoice no. DVT-AX-345678

Payment date: 03/12/2006

Reference	Designation	Qty	Unit price	Total CHF	Sales
Work					
SERVICE D	COMPLETE OVERHAUL	1	5500.00	5500.00	220
SERVICE D	REFRESHING COMPLETE CASE AND RHODIUM BATH	1	380.00	380.00	220
Exterior parts:					
JO.297.065.FP	FLAT GASKET	1	3.00	3.00	220
JO.197.075.FP	FLAT GASKET	1	4.00	4.00	220
JO.199.059.OS	FLAT ROUND GASKET	1	6.00	6.00	220
VI.261.036.BC	W.G.FIXATION SCREWS	10	4.00	40.00	220
AI.465.055.BC	WHITE GOLD "FOIL" PAIR OF HAND LENGTH: 10/13.50MM CALIBRE 2868	1	70.00	70.00	220
	SPECIAL DISCOUNT		-3003.00	-3003.00	
	Discount		-900.00	-900.00	
	Total CHF			2100.00	

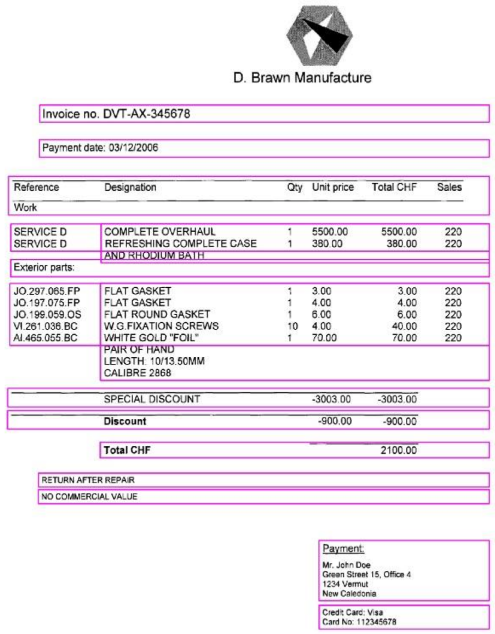
RETURN AFTER REPAIR

NO COMMERCIAL VALUE

Payment:

Mr. John Doe
Green Street 15, Office 4
1234 Vermut
New Caledonia

Credit Card: Visa
Card No: 112345678



D. Brawn Manufacture

Invoice no. DVT-AX-345678

Payment date: 03/12/2006

Reference	Designation	Qty	Unit price	Total CHF	Sales
Work					
SERVICE D	COMPLETE OVERHAUL	1	5500.00	5500.00	220
SERVICE D	REFRESHING COMPLETE CASE AND RHODIUM BATH	1	380.00	380.00	220
Exterior parts:					
JO.297.065.FP	FLAT GASKET	1	3.00	3.00	220
JO.197.075.FP	FLAT GASKET	1	4.00	4.00	220
JO.199.059.OS	FLAT ROUND GASKET	1	6.00	6.00	220
VI.261.036.BC	W.G.FIXATION SCREWS	10	4.00	40.00	220
AI.465.055.BC	WHITE GOLD "FOIL" PAIR OF HAND LENGTH: 10/13.50MM CALIBRE 2868	1	70.00	70.00	220
	SPECIAL DISCOUNT		-3003.00	-3003.00	
	Discount		-900.00	-900.00	
	Total CHF			2100.00	

RETURN AFTER REPAIR

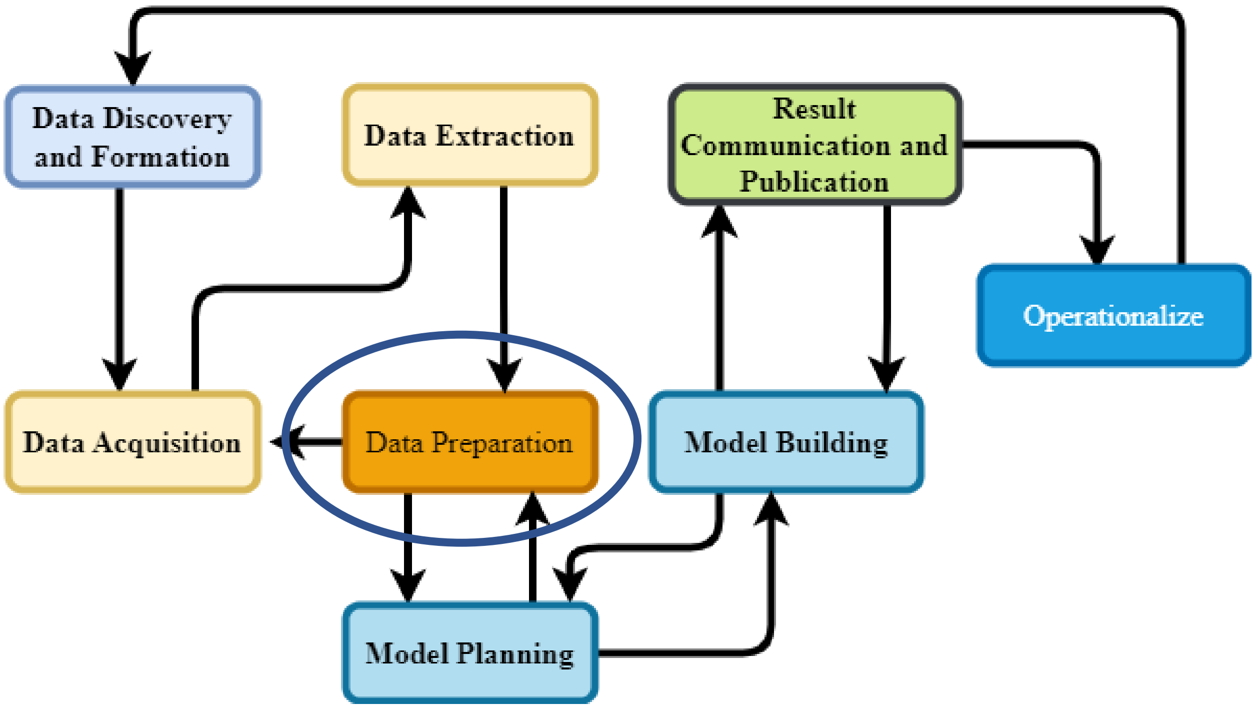
NO COMMERCIAL VALUE

Payment:

Mr. John Doe
Green Street 15, Office 4
1234 Vermut
New Caledonia

Credit Card: Visa
Card No: 112345678

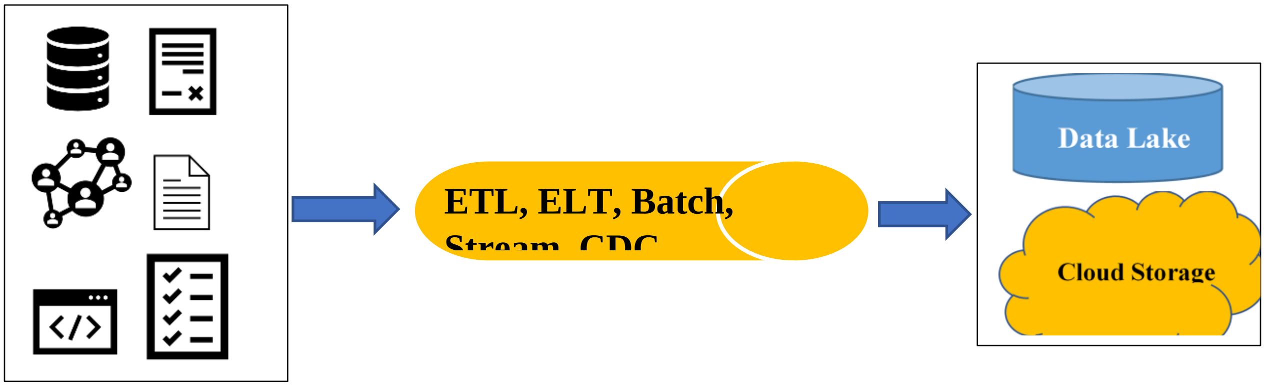
- ### Phase 4: Data Preparation



- Once data has been collected, it must be processed.
- It is the most crucial phase throughout the entire life cycle.
- Data preparation is the most time-consuming process:
    - About 50-80% of the total project duration
- It generally requires an analytical sandbox in which the team can analyze the data during the project. The data extraction is also part of the sandbox.
- The Analytical Sandbox is a standalone solution environment capable of capturing and processing large quantities of data from multiple sources in order to perform analytics in isolation the enterprise data warehouse.

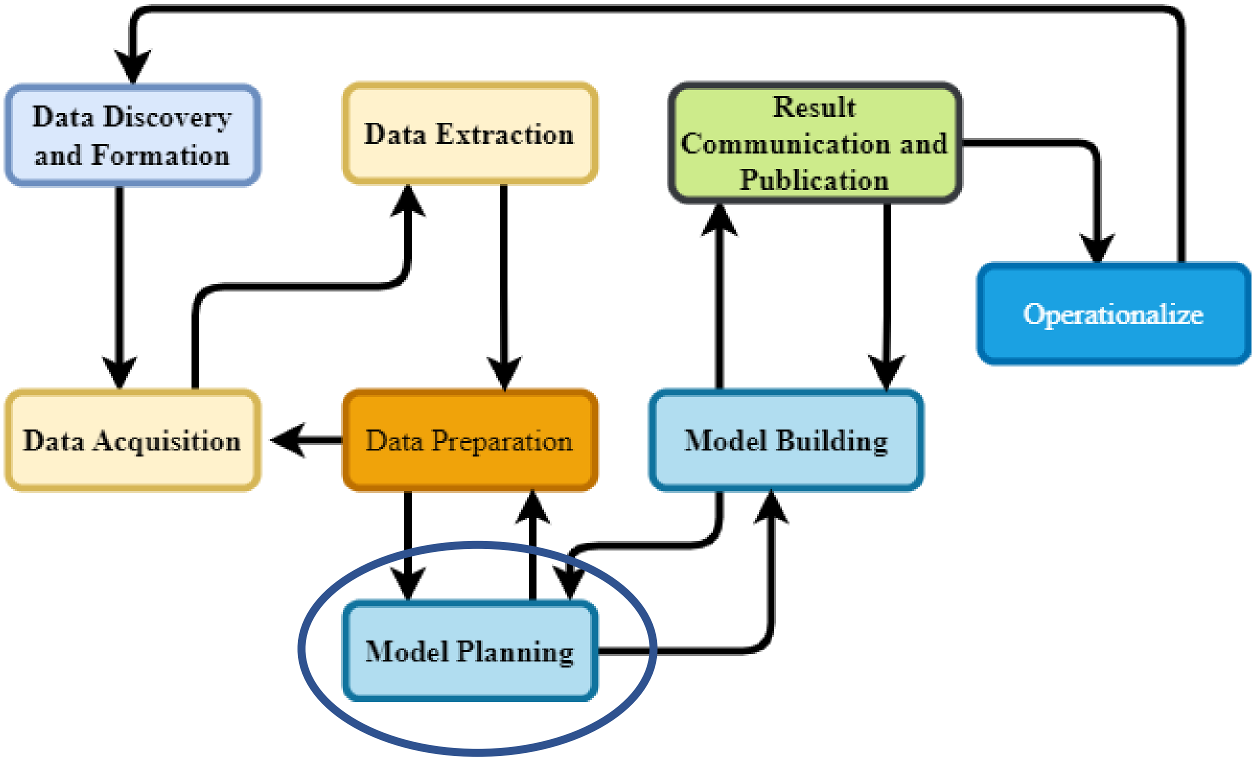
- A sandbox is a standalone developing environment used to capture and process large data from multiple sources for the duration of the project without affecting the original data and the application.
- Sandbox solution has been used to test application code and potential intrusion.
- There are many processes involved in cleaning, integrating, validating, combining multiple data sources, transforming, and preparing raw data for later analysis.
- Data preparation include the following:
  - Selecting relevant data,
  - Combining multiple data sources
  - Cleaning
  - Handling missing values
  - Handling incorrect data
  - Depending on your big data goals, you may want to check for outliers
  - Reducing your data dimensions:
    - Data may too big and you need to do some data reduction
  - Data Integration
    - Resolving any data conflicts and resolve data redundancy
    - Data coming from two different systems and both systems have table products.
  - Normalizing data:
    - Let us assume that we extracting data from three data sources that have different rating schema:

      - Rating from 1 to 5 with 1 being poor and 5 being excellent
      - Rating uses a “Positive”/” Negative”
      - Rating uses stars from one star to 5 stars.
- Different data preparation solutions:
  - Data wrangling aka data munging:
    - It is the process of removing errors and combining different data sources to make them easier to analyze.
    - It also consists of reorganizing, transforming, and mapping data from one "raw" form into more usable and formatted data for analysis
    - Functionalities:
      - Missing Data
      - Advanced Indexing
      - Duplicate Data
      - Grouping and Combining Data
      - Etc.
  - ELT/ETL: (The transformation part)



Change data capture (CDC) refers to the process of recognizing, tracking, and delivering changes in

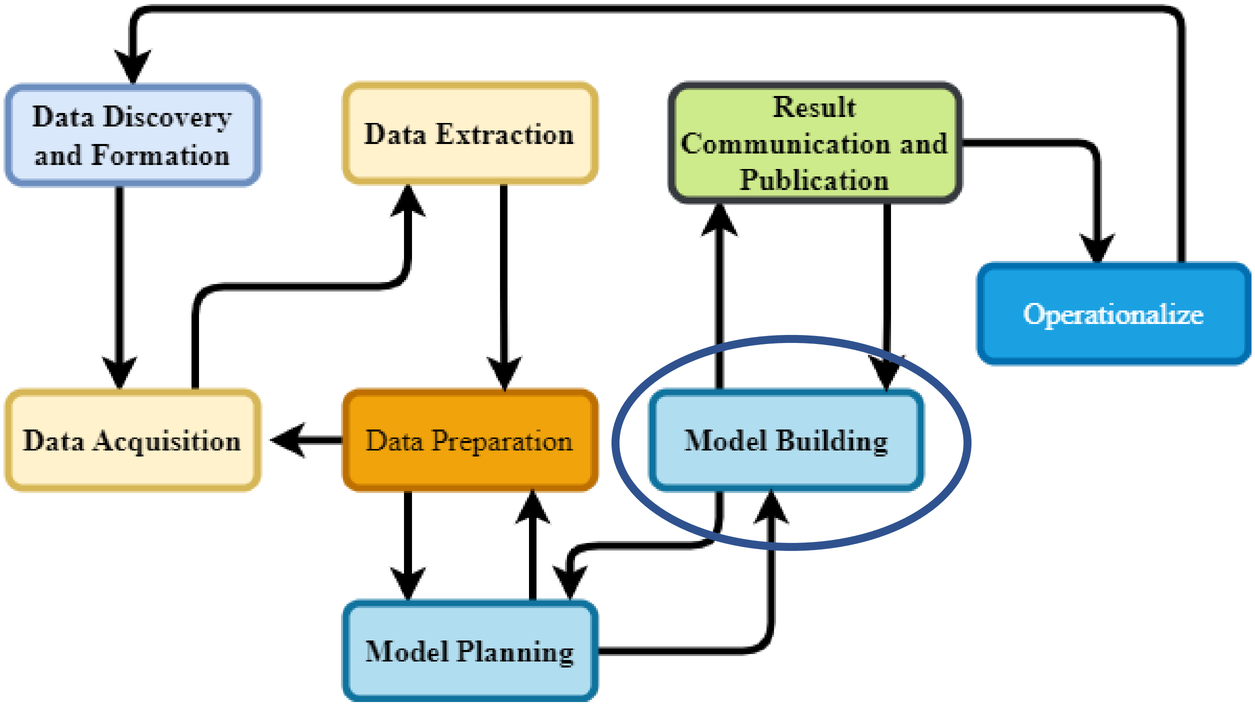
- ### Phase 5: Design a Model



- After preparing your data, your team need to start planning for analytical model using your project goals.
- You are planning for your big data analytics process(es).
  - Are you clustering, classifying, or needing a regression model?
- Identify tools that you are planning to use for your data analytics:
  - Your storage strategy: HDFS, NoSQL, data warehouse, etc.
  - Your programming language: R, Python, Scala, etc.
  - Your analytics solution
- Choose a data model to load your data to start your analytics:

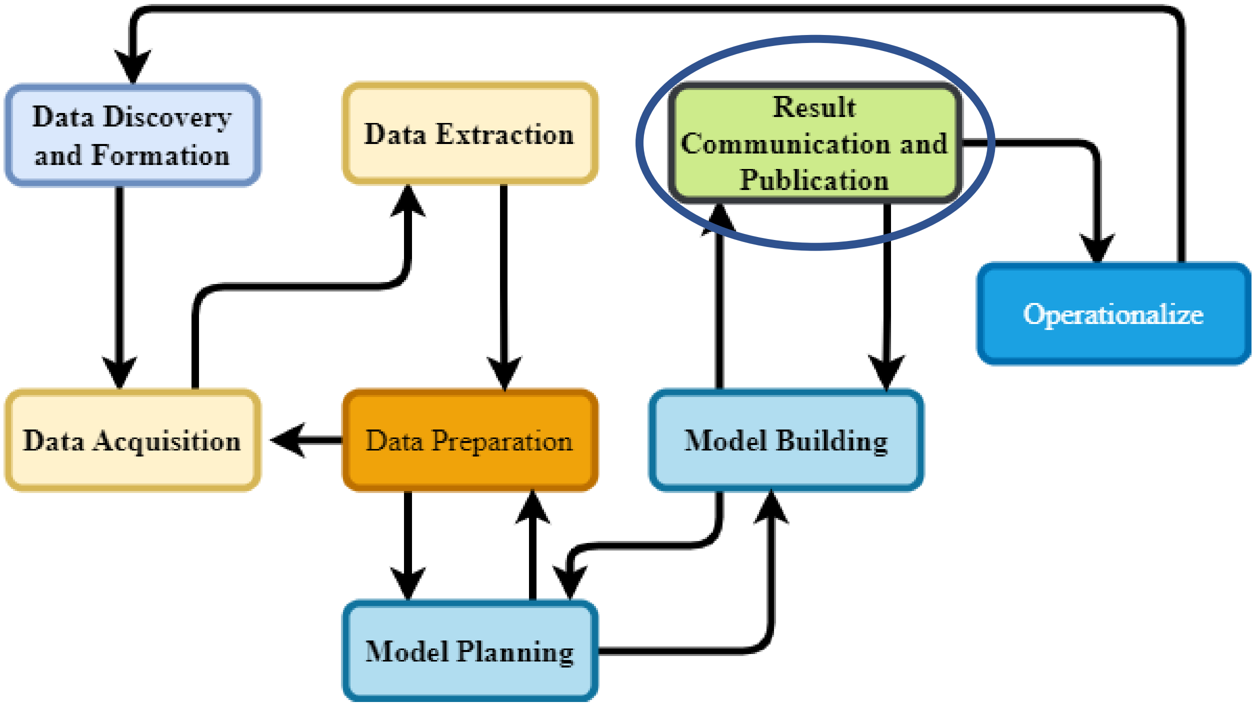
- ETL (Extract, Transform, and Load) transforms the data first using a set of business rules, before loading it into a sandbox.
- ELT (Extract, Load, and Transform) first loads raw data into the sandbox and then transform it.
- ETLT (Extract, Transform, Load, Transform) is a mixture; it has two transformation levels.

- **Phase 6: Model Building:**



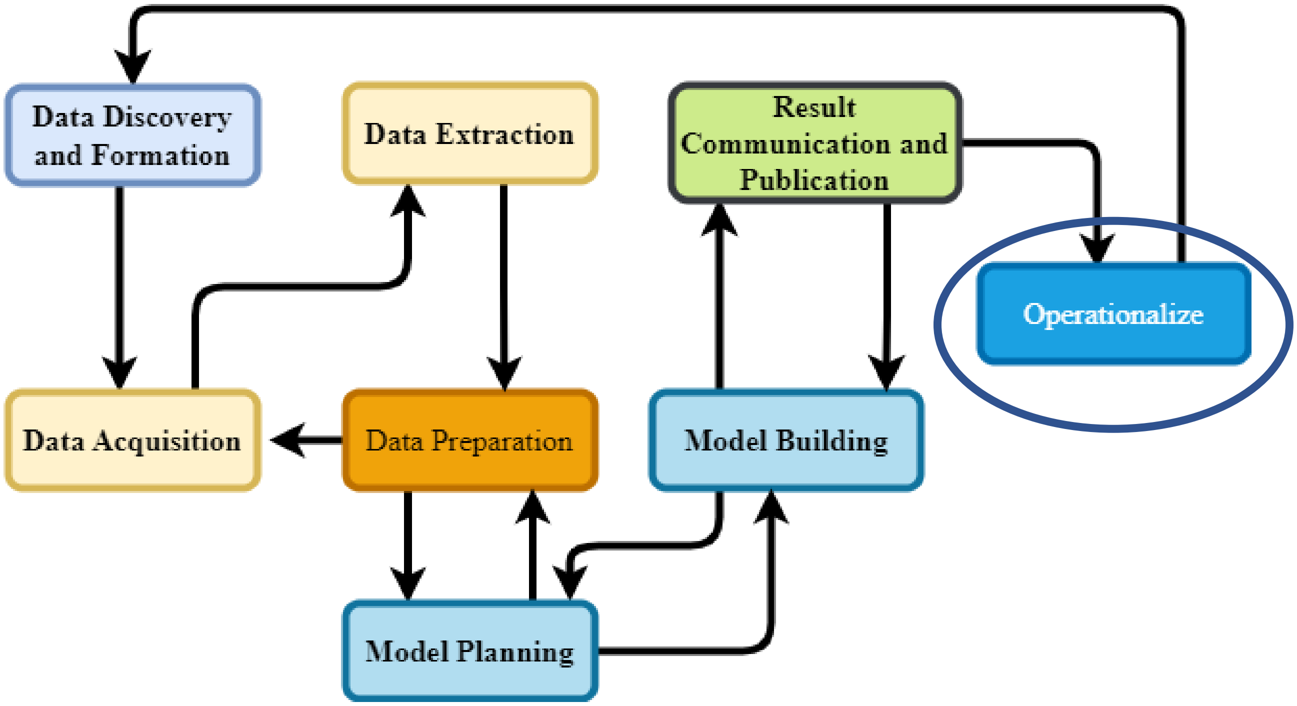
- Implement the model you design in Phase 5
- Depending on the objectives of your bid data project, you may need to use different implementation:
  - Regression analysis
  - Classification,
  - Clustering, etc.

- **Phase 7: Result Communication and Publication**



- Big data team summarize and present analysis results found in Phase 6 to stakeholders.
- Results are made available through dashboards and other big data tools.
- If the stakeholders are not satisfied and the process needs additional improvement, the team can go back to the model implement, design, etc.

- **Phase 8: Operationalize**



- In this phase, the team run the project in a controlled environment before broadening the work to full enterprise of users.
- This allows the team to measure the system performance and make adjustment before full deployment.
- The team delivers final reports, briefings, codes.