Big Data Analytics

Background and Overview of Data Analytics Lifecycle

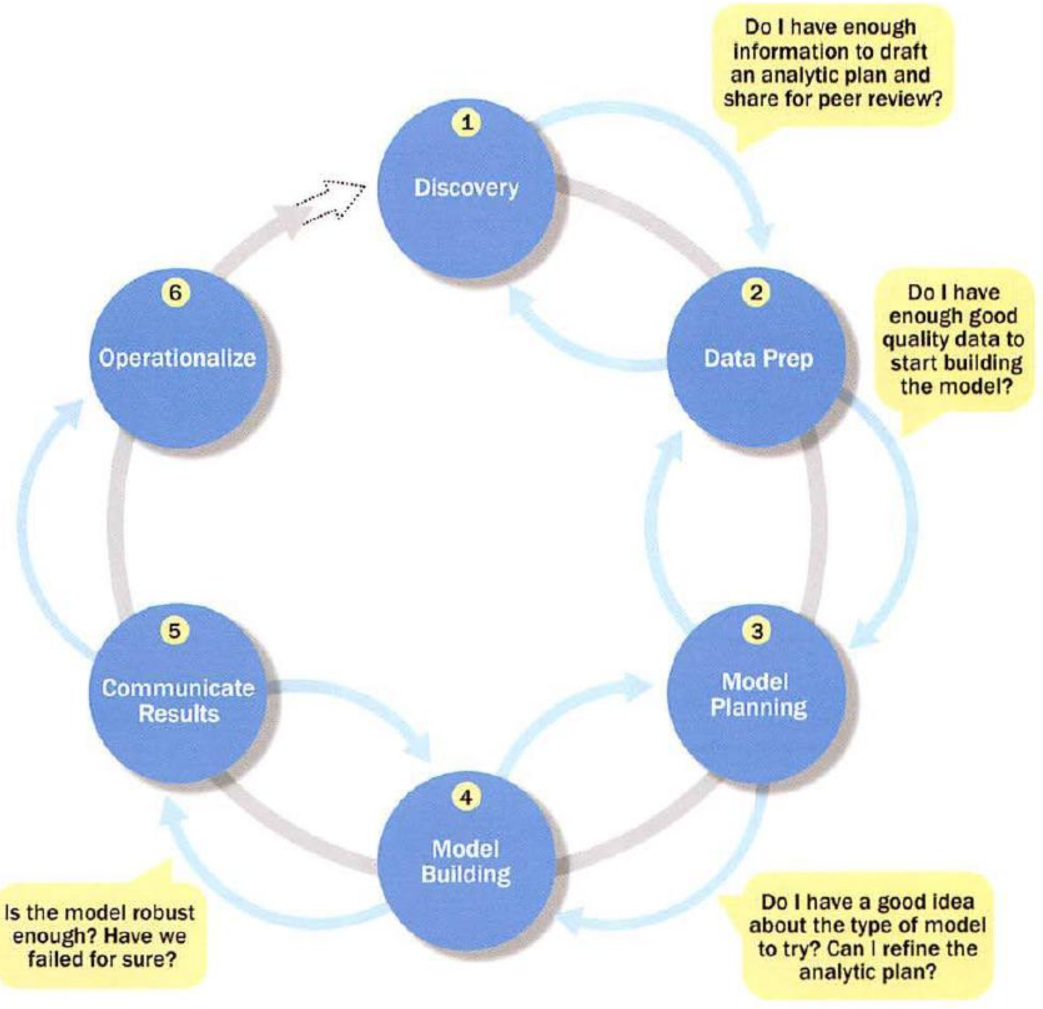

Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

FIGURE 2-2 *Overview of Data Analytics Lifecycle*

Phase 1 Discovery

- Learning the Business Domain
- Resources
- Framing the Problem
- Identifying Key Stakeholders
- Interviewing the Analytics Sponsor
- Developing Intial Hypothesis
- Identifying Potential Data Sources
    - Identify Data Sources
    - Capture Aggregate Data Sources
    - Review the Raw Data
    - Evaluate the Data Structures and Tools Needed
    - Scoop the type of Data Infrastructures needed for this type of Problem

Phase 2 Data Preparation

- Preparing the Analytic Sandbox
- Performing ETLT
- Learning about the Data
- Data Conditioning
- Survey and Visualization
- Common Tool for Data Preparation Phase
  - Hadoop, Alpine Miner, OpenRefine, Data Wrangler

Phase 3 Model Planning

- Data Exploration and Variable Selection
- Model Selection
- Common Tools for Model Planning Phase
  - R
  - SQL Analysis Services
  - SAS/ACCESS with data connectors like ODBC, JDBC and OLEDB

Phase 4 Model Building

- Does the model appear valid and accurate on the test data?
- Does the model output/behavior make sense to the domain experts? That is, does it appear as if the model is giving answers that make sense in this context?
- Do the parameter values of the fitted model make sense in the context of the domain?
- Is the model sufficiently accurate to meet the goal?
- Does the model avoid intolerable mistakes? Depending on context, false positives may be more serious or less serious than false negatives, for instance.
- Are more data or more inputs needed? Do any of the inputs need to be transformed or eliminated?
- Will the kind of model chosen support the runtime requirements?
- Is a different form of the model required to address the business problem? If so, go back to the model planning phase and revise the modeling approach.

Phase 4 Model Building…

- Common Tools for Module Building
  - Commercial Tools
    - SAS Enterprise Miner
    - SPSS Modeler
    - MATLAB
    - Alpine Miner
    - Statistica & Mathematica
  - Open Source Tools
    - R, PL/R
    - Octave
    - WEKA
    - Python
    - SQL – In-Database Implications such as MADLib

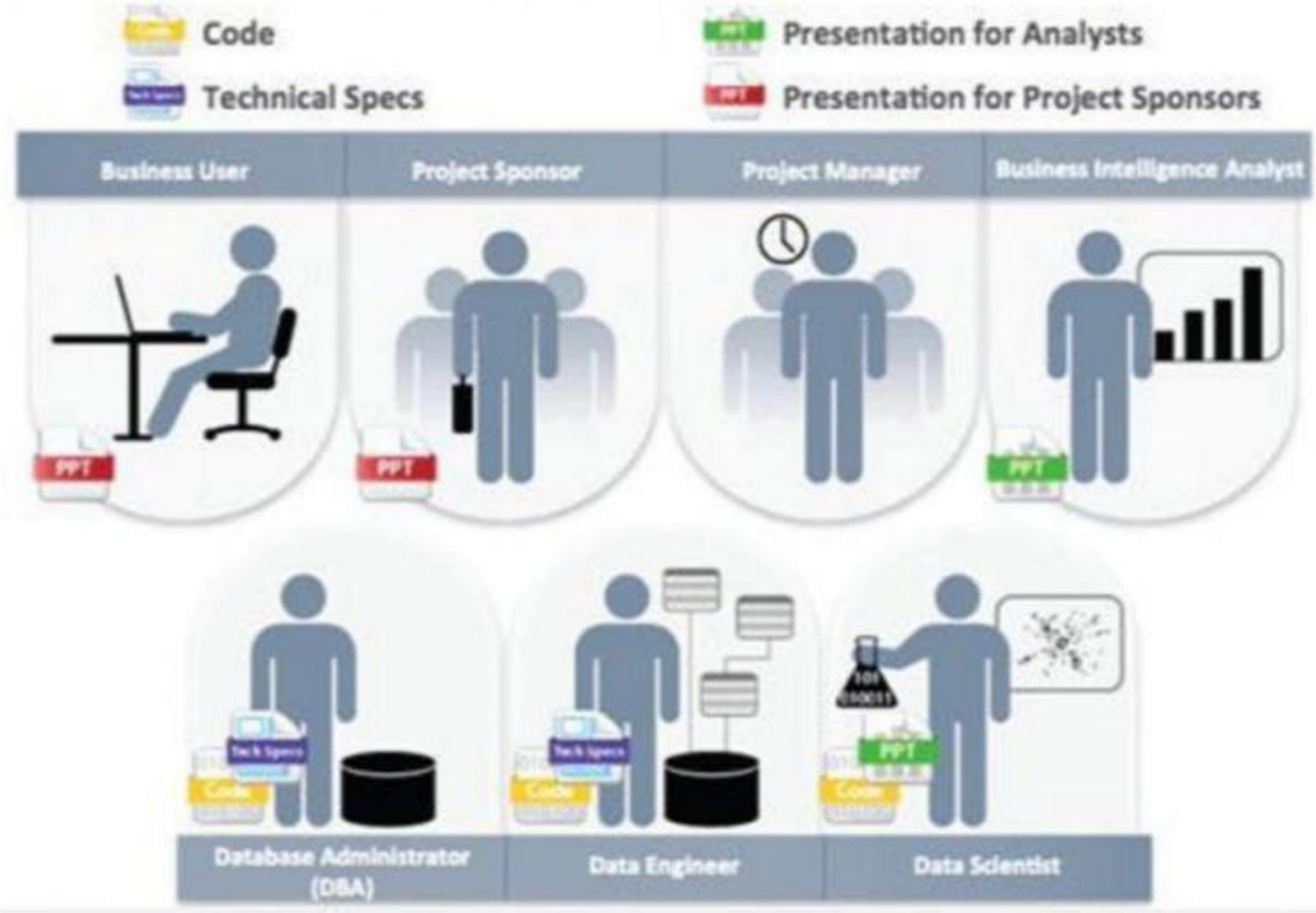
Phase 5 Communicate Results

- Sometimes teams have only done a superficial analysis, which is not robust enough to accept or reject a hypothesis.
- Other times, teams perform very robust analysis and are searching for ways to show results, even when results may not be there.
- It is important to strike a balance between these two extremes when it comes to analyzing data and being pragmatic in terms of showing real-world results.
- When conducting this assessment, determine if the results are statistically significant and valid.
- During this step, assess the results and identify which data points may have been surprising and which were in line with the hypotheses that were developed in Phase 1.
- As a result of this phase, the team will have documented the key findings and major insights derived from the analysis

Phase 6 Operationalize

- In the final phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way
- While scoping the effort involved in conducting a pilot project, consider running the model in a production environment for a discrete set of products or a single line of business, which tests the model in a live setting.
- This allows the team to learn from the deployment and make any needed adjustments before launching the model across the enterprise.
- Part of the operationalizing phase includes creating a mechanism for performing ongoing monitoring of model accuracy and.
- if accuracy degrades, finding ways to retrain the model.
- Refer Fig. on next slide

Key Outputs from a Successful Analytics Project



The Age of the Data Product

Unit II Syllabus

- Data Product, Building Data Products at Scale with Hadoop, Data Science Pipeline and Hadoop Ecosystem

  Ref: (Chapter 1 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

- Operating System for Big Data: Concepts, Hadoop Architecture, Working with Distributed file system, Working with Distributed Computation,

  Ref: (Chapter 2 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

- Framework for Python and Hadoop Streaming, Hadoop Streaming, MapReduce with Python, Advanced MapReduce.

  Ref: (Chapter 3 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

- In-Memory Computing with Spark, Spark Basics, Interactive Spark with PySpark, Writing Spark Applications

  Ref: (Chapter 4 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

Today’s Topics

- Data Product
- Building Data Products at Scale with Hadoop
- Data Science Pipeline and Hadoop Ecosystem

Ref: (Chapter 1 - Data Analytics with Hadoop By Benjamin Bengfort & Jenny Kim)

The Age of the Data Product

- The information revolution had a transformative effect on society, academia, and business.

- The revolution of networked communication systems and the Internet has created a surplus of a valuable new material—data—and transformed us all into both consumers and producers.

- Data increasingly affects every aspect of our lives, from the food we eat, to our social interactions, to the way we work and play.

- In turn, we have developed a reasonable expectation for products and services that are highly personalized and finely tuned to our bodies, our lives, and our businesses, creating a market for a new information technology—the data product.

The Age of the Data Product…

- Datasets + ML Algorithms led to immediate and Novel results.
- The Buzzword trends surrounding ‘Big Data’ are related to the seemingly inexhaustible innovation that is available due to the large number of models and data sources.
- Data products are created with data science workflows, specifically through the application of models, usually predictive or inferential, to a domain-specific dataset.
- While the potential for innovation is great, the scientific or experimental mindset that is required to discover data sources and correctly model or mine patterns is not typically taught to programmers or analysts.
- The PhD’s holders (Who are trained for analytics and doing experiments) with some programming skills immediately lead to the Data Science expertise.
- All of us can’t be a PhD. So We need some data products to work with Big Data.

What is a Data Product

Data Product is any application that combines data and algorithms.

Data product is the combination of data with statistical algorithms that are used for inference or prediction. Ex: Amazon Recommendations

A data product is an economic engine.
It derives value from data and then produces more data, more value, in return.

A data application acquires its value from the data itself, and creates more data as a result. It’s not just an application with data; it’s a data product.
– Mike Loukides

Data products are systems that learn from data, are self-adapting, and are broadly applicable. Ex: Nest Thermostat

Building Data Products at Scale with Hadoop

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

– Josh Wills

Data Products are built at the intersection of data, domain knowledge, software engineering, and analytics.

- Harlan Harris

- Domain knowledge and analytics are the tools used to build the data engine, usually via experimentation, hence the “science” part of data science.

Typical analytical workflow:

Ingestion → Wrangling → Modeling → Reporting and Visualization

Leveraging Large Datasets

- The cognitive process of making sense of data involves high-level overviews of data, zooming into specified levels of detail, and moving back out again.

- More data can be both tightly tuned patterns and signals just as much as it can be noise and distractions.

- Statistical methodologies give us the means to deal with simultaneously noisy and meaningful data, either by describing the data through aggregations and indices or inferentially by directly modeling the data.

- Statistical techniques that attempt to take into account rare events leverage a computer's power to track multiple data points simultaneously, but require more computing resources.

- Statistical methods have traditionally taken a sampling approach to much larger datasets, wherein a smaller subset of the data is used as an estimated stand-in for the entire population

Leveraging Large Datasets…

- The past decade has seen the unprecedented rise of data science, fueled by the seemingly limitless combination of data and machine learning algorithms to produce truly novel results.
- Scale comes not just from the amount of data, but from the number of facets that exploration requires —a forest view for individual trees.

- Hadoop, an open source implementation of two papers written at Google that describe a complete distributed computing system, caused the age of big data.
- Data warehouse systems as computationally powerful as Hadoop predate those papers in both industry and academia.
- What makes Hadoop different is partly the economics of data processing and partly the fact that Hadoop is a platform.

- Hadoop was released right at the moment when technology needed a solution to do data analytics at scale, not just for population-level statistics, but also for individual generalizability and insight

Hadoop for Data Products

- Hadoop comes from big companies with big data challenges like Google, Facebook, and Yahoo.
- Data challenges are resolved for tech giants, Commercial and governmental entities from large to small: enterprises to startups, federal agencies to cities, and even individuals.
- Computing resources are also becoming ubiquitous and cheap.
- Cloud computing resources such as Amazon EC2 and Google Compute Engine mean that data scientists have unprecedented on-demand, instant access to large-scale clusters for relatively little money and no data center management.
- Hadoop has made big data computing democratic and accessible.

Hadoop for Data Products…

- Examples
- In 2011, Lady Gaga released her album Born This Way, an event that was broadcast by approximately 1.3 trillion social media impressions from “likes” to tweets to images and videos.
- More recently, in 2015, the New York City Police Department installed a $1.5 million dollar acoustic sensor network called ShotSpotter.
- As of 2012, the Affordable Care Act man- dates that health plans implement standardized secure and confidential electronic exchange of health records.
- Connected homes and mobile devices, along with other personal sensors, are generating huge amounts of individual data, which among other things sparks concern about privacy.
- In 2015, researchers in the United Kingdom created the Hub of All Things (HAT)

Hadoop for Data Products…

- Large-scale, individual data analytics have traditionally been the realm of social net- works like Facebook and Twitter.
- Cities deal with unique data challenges, but whereas the generalization of a typical city could suffice for many analytics, new data challenges are arising that must be explored on a per-city basis.
- How do technologies provide value to consumers utilizing their personal health records without aggregation to others because of privacy issues? Can we make personal data mining for medical diagnosis secure?
- In order to answer these questions on a routine and meaningful (individual) basis, a data product is required. Applications like Place, ShotSpotter, quantified self products, and HAT derive their value from data and generate new data by providing an application platform and decision-making resources for people to act upon.
- Big data workflows and Hadoop have made these applications possible and personalized.

The Data Science Pipeline and the Hadoop Ecosystem

In each phase, an analyst transforms an initial dataset, augmenting or ingesting it from a variety of data sources, wrangling it into a normal form that can be computed upon, either with descriptive or inferential statistical methods, before producing a result via visualization or reporting mechanisms

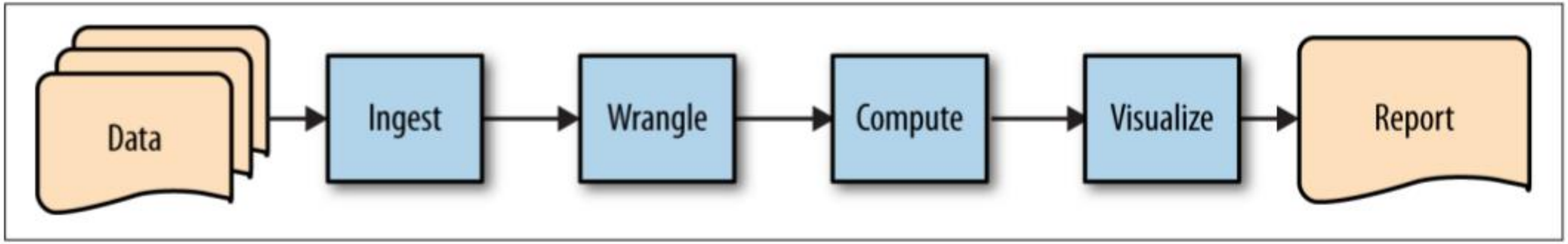


Figure 1-1. The data science pipeline

The Data Science Pipeline and the Hadoop Ecosystem

- These analytical procedures are usually designed to answer specific questions, or to investigate the relationship of data to some business practice for validation or decision making.

- Original workflow model has driven most early data science thought.
- This workflow is intended to produce something that allows humans to make decisions.
- This human-powered model is not a scalable solution in the face of exponential growth in the volume and velocity of data that many organizations are now grappling with.
- In addition to the limitations of scale, the human-centric and one-way design of this workflow precludes the ability to efficiently design self-adapting systems that are able to learn.

- To create a framework that allows the construction of scalable, automated solutions to interpret data and generate insights, we must revise the data science pipeline into a framework that incorporates a feedback loop for machine learning methods.

Big Data Workflows

- Iterative Model

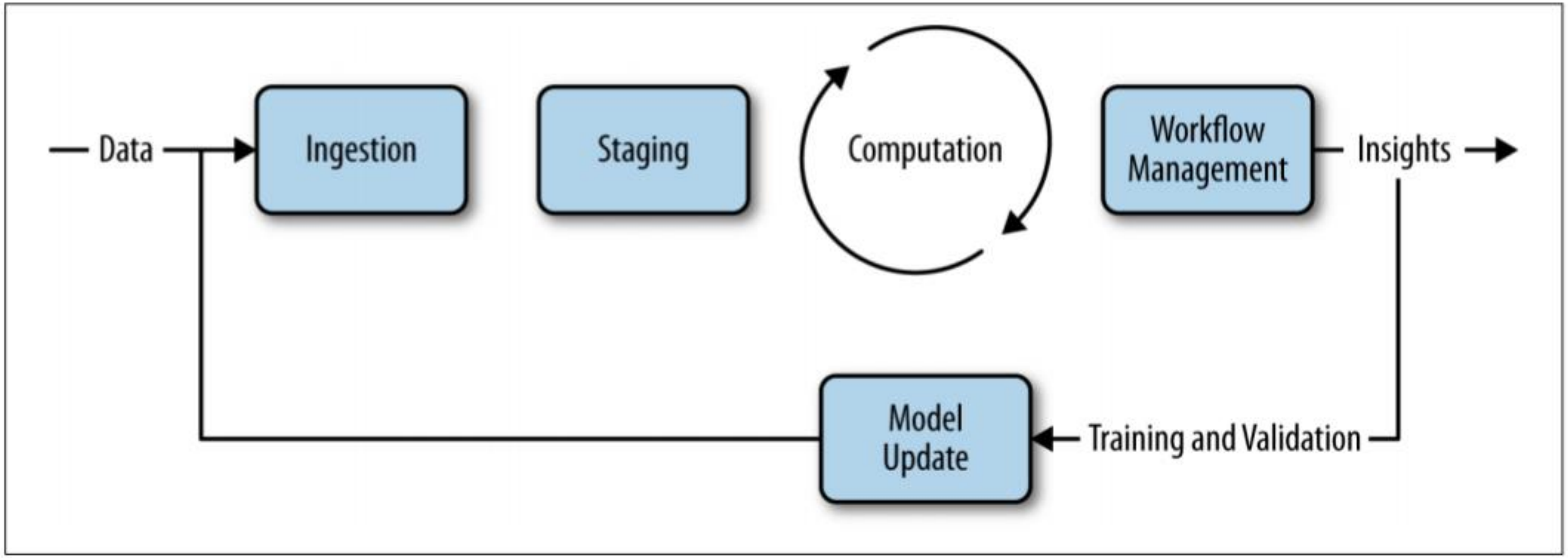


Figure 1-2. The big data pipeline

Big Data Workflows

- This model in its simplest form takes raw data and converts it into insights.
- The crucial distinction, however, is that the data product pipeline builds in the step to operationalize and automate the workflow.
- By converting the ingestion, staging, and computation steps into an automated workflow, this step ultimately produces a reusable data product as the output.
- The workflow management step also introduces a feedback flow mechanism, where the output from one job execution can be automatically fed in as the data input for the next iteration, and thus provides the necessary self-adapting framework for machine learning applications.

Big Data Workflows…

- Hadoop has specifically evolved into an ecosystem of tools that operationalize some part of this pipeline.
- For example, Sqoop and Kafka are designed for ingestion, allow- ing the import of relational databases into Hadoop or distributed message queues for on-demand processing.
- In Hadoop, data warehouses such as Hive and HBase provide data management opportunities at scale.
- Libraries such as Spark's GraphX and MLlib or Mahout provide analytical packages for largescale computation as well as validation.

USERS
MR. CLIENT MACHINE

Thank You….

Revise the topics from Syllabus References…

Fill Your Attendance Form …..!

Syllabus References

1. Big Data and Analytics, Subhashini Chellappan Seema Acharya, Wiley
2. Data Analytics with Hadoop *An Introduction for Data Scientists*, Benjamin Bengfort and Jenny Kim, O'Reilly
3. Big Data and Hadoop, V.K Jain, Khanna Publishing

   https://books.google.co.in/books?id=i6NODQAAQBAJ&pg=PA122&source=gbs_toc_r&cad=4#v=onepage&q&f=true