

# AJAY KUMAR GARG ENGINEERING COLLEGE, GHAZIABAD



Computer Science And Engineering

**Introduction to Data Analytics and Visualization**

## **Introduction to Data**

Presented By:

**Ms. Neeharika Tripathi**

Assistant Professor

Department of Computer Science And Engineering

# Data Science Vs Data Analytics

- Data Science is a field that deals with extracting meaningful information and insights by applying various algorithms preprocessing and scientific methods on structured and unstructured data.
- Data Analytics is used to get conclusions by processing the raw data. It is helpful in various businesses as it helps the company to make decisions based on the conclusions from the data.

# Need of Data Analytics

- It helps organizations to make sense of this data, turning it into actionable insights. These insights can be used to improve products and services, enhance experiences, streamline operations, and increase profitability.

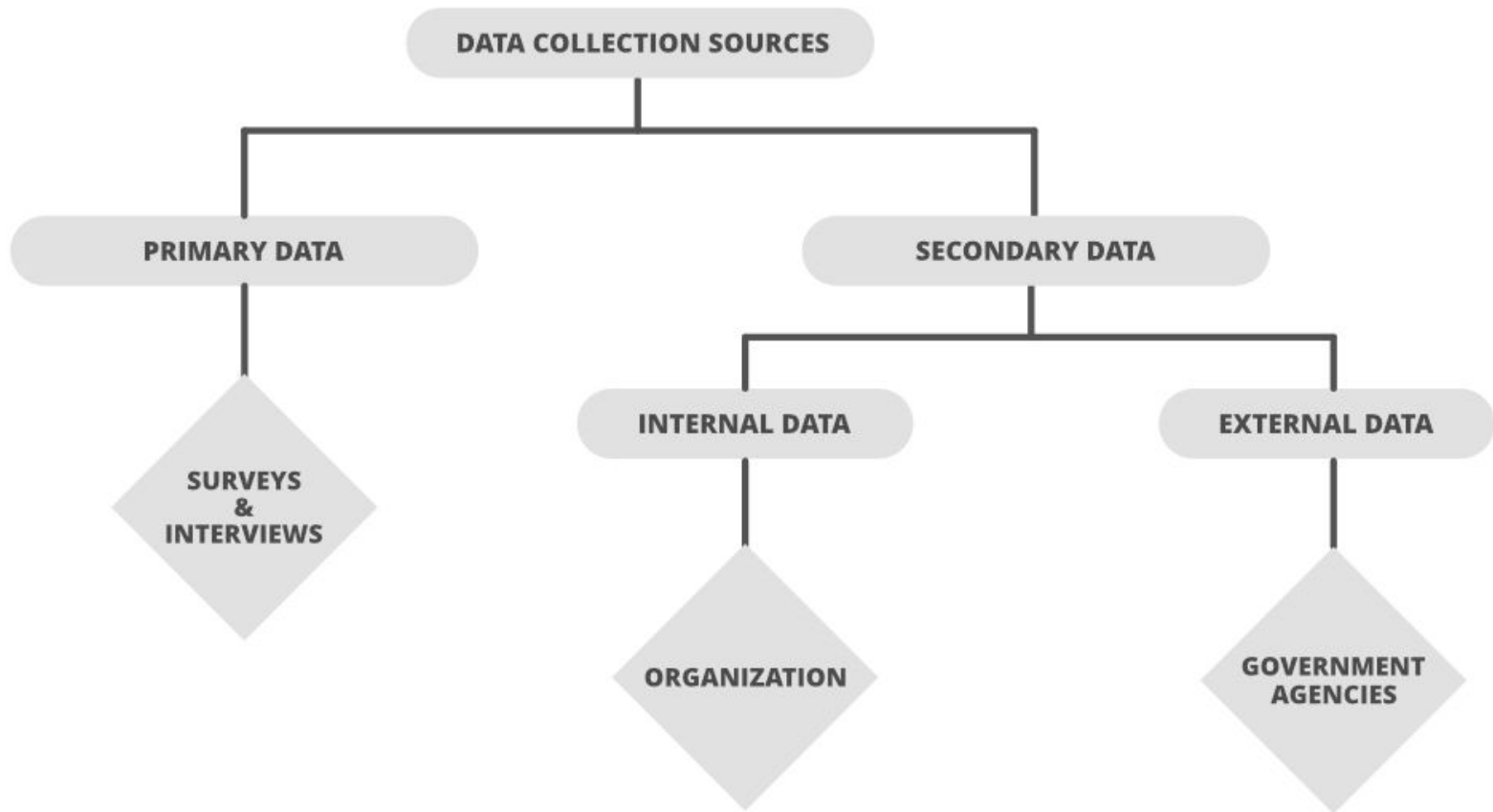
# Data

- Data is a collection of raw, unorganized facts and details like text, observations, figures, symbols and descriptions of things etc.
- Data is measured in terms of bits and bytes – which are basic units of information in the context of computer storage and processing.

# Data Vs Information

Data	Information
Data is unorganised and unrefined facts	Information comprises processed, organised data presented in a meaningful context
Data is an individual unit that contains raw materials which do not carry any specific meaning.	Information is a group of data that collectively carries a logical meaning.
Data doesn't depend on information.	Information depends on data.
Raw data alone is insufficient for decision making	Information is sufficient for decision making
An example of data is a student's test score	The average score of a class is the information derived from the given data.

# Sources of Data





# Data Classification

- Data
  - Structured Data
  - Unstructured Data
  - Semi- Structured Data

# Structured Data

- Structured data is created using a fixed schema and is maintained in tabular format.
- **Examples –**
- Relational data, Geo-location, credit card numbers, addresses, etc.



# Unstructured Data

- It is defined as the data in which is not follow a pre-defined standard or you can say that any does not follow any organized format.
- **Examples –**
- Word, PDF, text, media logs, etc.

# Semi-Structured Data

- Semi-structured data is information that does not reside in a relational database but that have some organizational properties that make it easier to analyze.
- [XML data](#), Email message data

# Characteristics of Data



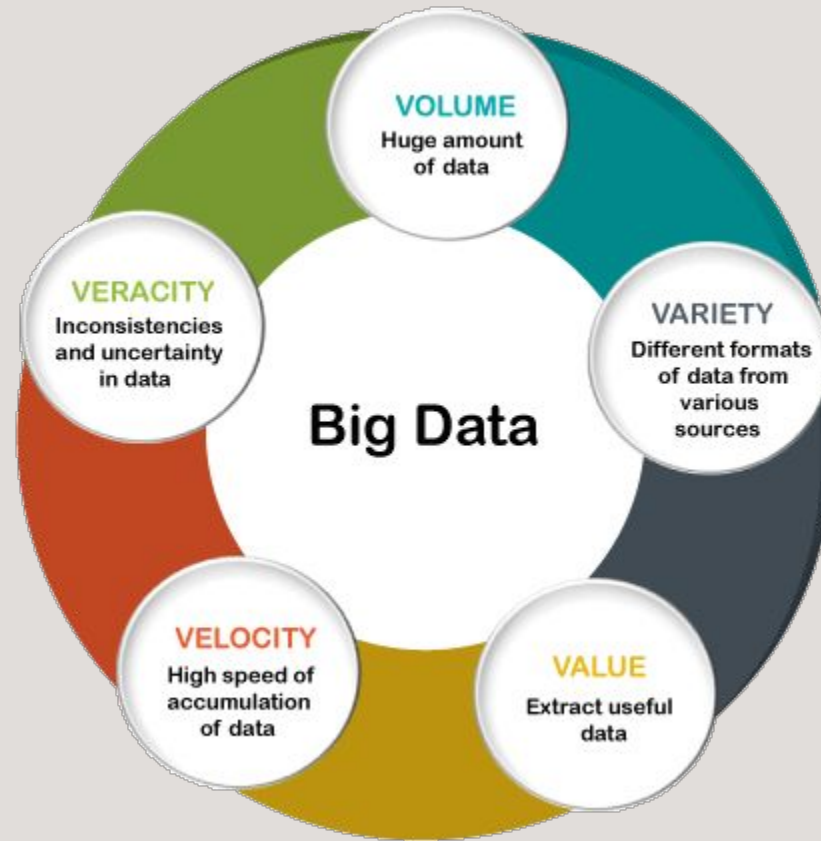
# Introduction to Big Data Platform

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time.
- It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.

# Examples of Big data

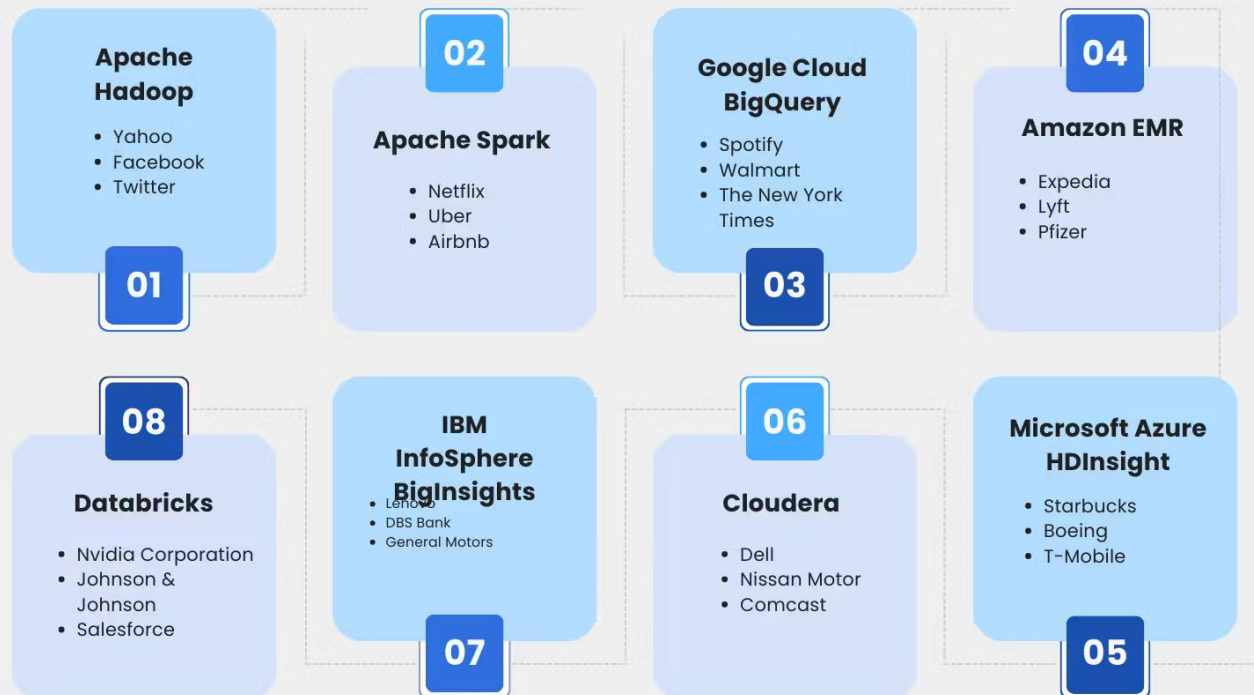
- **Social Media:** The statistic shows that **500+terabytes** of new data get ingested into the databases of social media site **Facebook**, every day.
- A single **Jet engine** can generate **10+terabytes** of data in **30 minutes** of flight time. With many thousand flights per day, generation of data reaches up to many **Petabytes**.

# Characteristics Of Big Data



# Big Data Platforms

## The Best Big Data Platforms



# Apache Hadoop

- It is an open-source framework that enables distributed processing for massive datasets throughout clusters.
- Hadoop provides a scalable and cost-effective solution for storing, processing, and analyzing massive amounts of structured and unstructured data.
- The prominent companies that use Apache Hadoop are: Yahoo, Facebook, Twitter



# Apache Spark

- Apache Spark is a unified analytics engine for batch processing, streaming data, machine learning, and graph processing.
- It is designed to perform data processing tasks in-memory and achieve significantly faster processing times than traditional disk-based systems.
- Spark also supports various programming languages, including Java, Scala, [Python, and R](#), making it accessible to a wide range of developers. The prominent companies that use Apache Spark include: Netflix, Uber, Airbnb

# Google Cloud BigQuery

- Google Cloud BigQuery is a top-rated big data platform that provides a fully managed and serverless data warehouse solution.
- BigQuery is designed to handle petabytes of data and allows users to run SQL queries on large datasets with impressive speed and efficiency.
- The prominent companies that use Google Cloud BigQuery are: Spotify, Walmart, The New York Times

# Amazon EMR

- EMR allows users to quickly provision and manage clusters of virtual servers, known as instances, to process data in parallel.
- EMR integrates seamlessly with other AWS services, such as Amazon S3 for data storage and Amazon Redshift for data warehousing, enabling a comprehensive big data ecosystem.
- The prominent companies that use Amazon EMR are: Expedia, Lyft, Pfizer

# Microsoft Azure HDInsight

- Provides a fully managed cloud service for processing and analyzing large datasets using popular open-source frameworks such as Apache Hadoop, Apache Spark, Apache Hive, and Apache HBase.
- HDInsight integrates seamlessly with other Azure services, such as Azure Data Lake Storage and Azure Synapse Analytics, offering a comprehensive ecosystem of Microsoft Azure services.
- The prominent companies that use Microsoft Azure HDInsight are: Starbucks, Boeing, T-Mobile

# Cloudera

- Cloudera is a leading big data platform that offers a comprehensive suite of tools and services designed to help organizations effectively manage and analyze large volumes of data.
- Cloudera offers a unified platform that integrates various components such as Hadoop Distributed File System (HDFS), Apache Spark, and Apache Hive, enabling users to perform various data processing and analytics tasks.
- The prominent companies that use Cloudera are: Dell, Nissan Motor, Comcast.

# IBM InfoSphere BigInsights


- IBM InfoSphere BigInsights provides a user-friendly interface and intuitive data exploration and visualization tools.
- BigInsights is built on top of Apache Hadoop and Apache Spark, and it integrates with other IBM products and services, such as IBM DB2, IBM SPSS Modeler, and IBM Watson Analytics. The prominent companies that use IBM InfoSphere BigInsights are: Lenovo, DBS Bank, General Motors

# Databricks

- Databricks is a prominent big data platform built on Apache Spark.
- Databricks simplifies the process of building and deploying big data applications by providing a scalable and fully managed infrastructure. It allows users to process large datasets in real-time, perform complex analytics, and build machine learning models using Spark's powerful capabilities.
- The prominent companies that use Databricks are: Nvidia Corporation, Johnson & Johnson, Salesforce



# Need of data analytics

- 
- ❖ 1. Improved Decision Making
  - ❖ 2. Better Customer Service
  - ❖ 3. Efficient Operations
  - ❖ 4. Effective Marketing



# Types of data Analytics

- ❖ **Descriptive:** Descriptive analytics is when you assess historical data and try to identify specific patterns. The main goal is to answer what happened and if it was expected or not, making comparisons with other timeframes.
- ❖ **Diagnostic:** When we know what's going on, the next step is to understand why. So you may have performed some descriptive analytics techniques and you were able to identify that sales went up by 12%. Diagnostic analytics is there to help identify why this happened and what actually worked for your business.
- ❖ **Predictive:** Predictive analytics involves sophisticated techniques that can help you use the patterns observed and make forecasts about future performance, e.g., [financial data analytics](#). While this may require specific expertise, it's extremely useful in order to be better prepared for the future.
- ❖ **Prescriptive:** Last but not least, prescriptive analytics techniques can help you identify the best course of action. This type of analytics is frequently used by marketers to draft their strategies and achieve better results.

# Modern data analytic tools

## ❖ **Apache Hadoop:**

It's a Java-based open-source platform that is being used to store and process big data. It is built on a cluster system that allows the system to process data efficiently and let the data run parallel. It can process both structured and unstructured data from one server to multiple computers.

## ❖ **Cassandra**

APACHE Cassandra is an open-source NoSQL distributed database that is used to fetch large amounts of data. It is capable of delivering thousands of operations every second and can handle petabytes of resources with almost zero downtime.

# Modern data analytic tools

## ❖ **SAS (Statistical Analytical System):**

Statistical Analytical System or SAS allows a user to access the data in any format (SAS tables or Excel worksheets). Besides that it also offers a cloud platform for business analytics called **SAS Viya** and also to get a strong grip on AI & ML, they have introduced new tools and products.

## ❖ **Spark**

APACHE Spark is another framework that is used to process data and perform numerous tasks on a large scale. It is also used to process data via multiple computers with the help of distributing tools.

# Modern data analytic tools

## ❖ **Apache Storm**

A storm is a robust, user-friendly tool used for data analytics, especially in small companies. The best part about the storm is that it has no language barrier (programming) in it and can support any of them. It was designed to handle a pool of large data in fault-tolerance and horizontally scalable methods.

## **4. Rapid Miner**

It's a fully automated visual workflow design tool used for data analytics. It's a no-code platform and users aren't required to code for segregating data. Today, it is being heavily used in many industries such as ed-tech, training, research, etc.

# Applications of data analytics

- ❖ Transportation
- ❖ Security
- ❖ Internet Web search results
- ❖ Marketing and digital advertising

# Analytics and Reporting

**Analytics:** Analytics is about diving deeper into your data and reports in order to look for insights. It's actually an attempt to answer **why** something is happening. Analytics powers up decision-making as the main goal is to make sense of the data explaining the reason behind the reported numbers.

**Reporting:** Data reporting is about taking the available information (e.g. your dataset), organizing it, and displaying it in a well-structured and digestible format we call “reports”.

# Analytics vs reporting

## ❖ Purpose:

The purpose of reports is to take data and organize it into clear information. Analytics aims to take that data and provide insights that drive better business decisions.

## ❖ Methods:

When discussing reports or reporting, you may use language such as organizing, formatting, building, configuring, consolidating, or summarizing. Analytics employs words and phrases like investigating, performing a “deep dive,” questioning, examining, interpreting, comparing, and confirming



# Analytics vs reporting

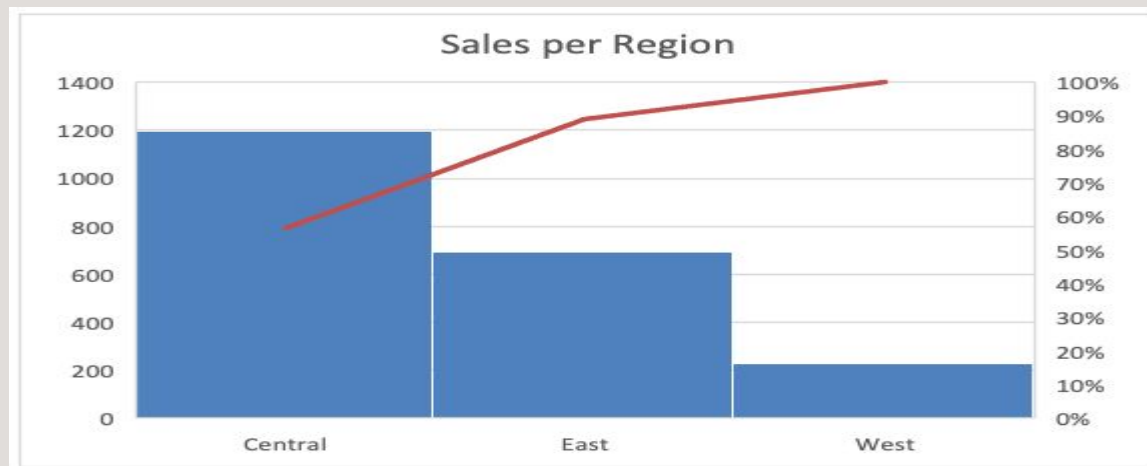
## ❖ Value:

Data analytics transforms data into information whereas reporting transforms the information into insights & recommendations.

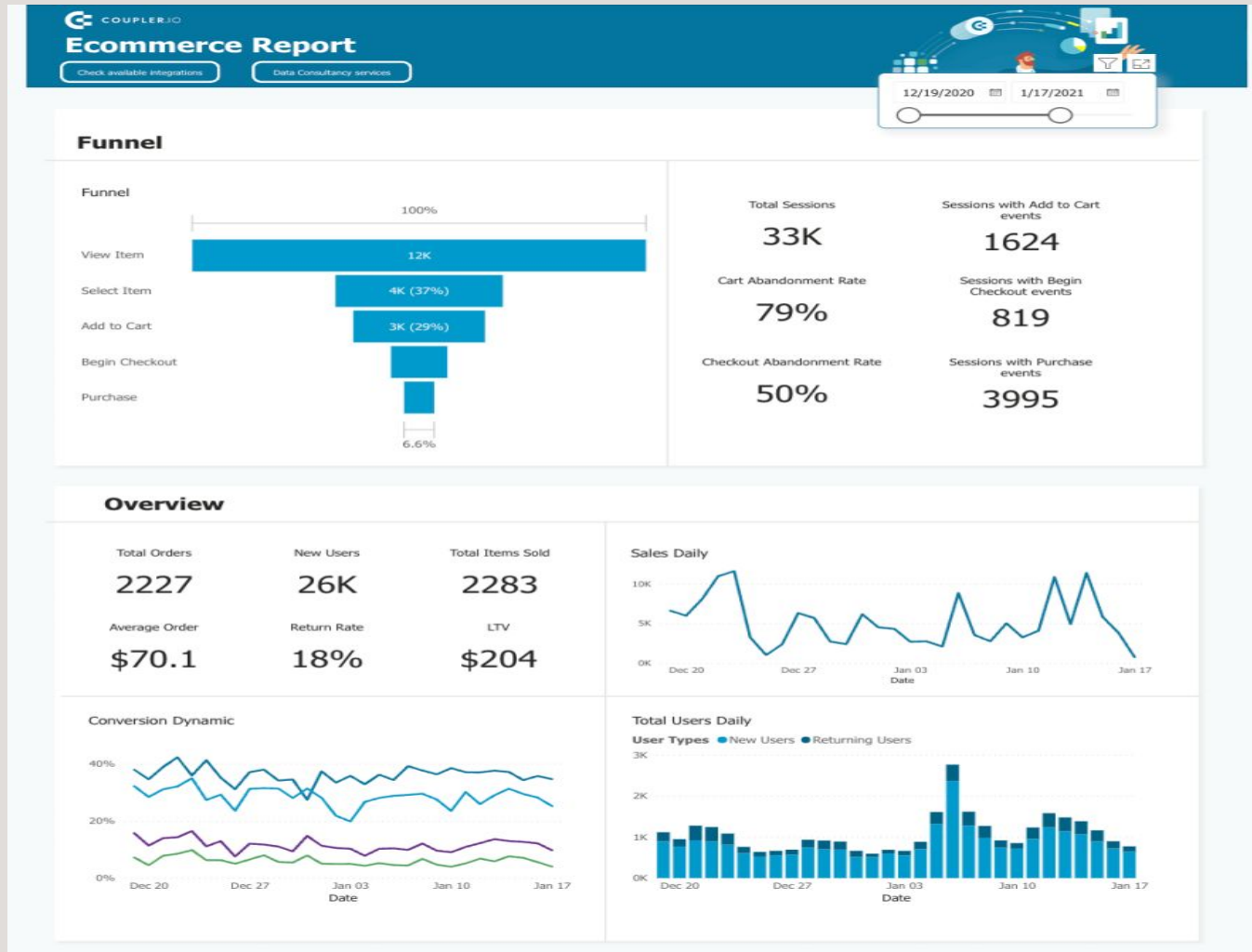


# Examples of Reports

	A	B	C	D	E	F	G
1	OrderDate	Region	Rep	Item	Units	Unit Cost	Total
2	1/6/21	East	Jones	Pencil	95	1.99	189.05
3	1/23/21	Central	Kivell	Binder	50	19.99	999.50
4	2/9/21	Central	Jardine	Pencil	36	4.99	179.64
5	2/26/21	Central	Gill	Pen	27	19.99	539.73
6	3/15/21	West	Sorvino	Pencil	56	2.99	167.44
7	4/1/21	East	Jones	Binder	60	4.99	299.40
8	4/18/21	Central	Andrews	Pencil	75	1.99	149.25
9	5/5/21	Central	Jardine	Pencil	90	4.99	449.10
10	5/22/21	West	Thompson	Pencil	32	1.99	63.68
11	6/8/21	East	Jones	Binder	60	8.99	539.40
12	6/25/21	Central	Morgan	Pencil	90	4.99	449.10
13	7/12/21	East	Howard	Binder	29	1.99	57.71
14	7/29/21	East	Parent	Binder	81	19.99	1619.19
15	8/15/21	East	Jones	Pencil	35	4.99	174.65
16	9/1/21	Central	Smith	Desk	2	125.00	250.00
17	9/18/21	East	Jones	Pen Set	16	15.99	255.84
18	10/5/21	Central	Morgan	Binder	28	8.99	251.72
19	10/22/21	East	Jones	Pen	64	8.99	575.36
20	11/8/21	East	Parent	Pen	15	19.99	299.85



# Analysis



# Evolution of Analytic Scalability

Traditional analytics collects data from heterogeneous data sources and we had to pull all data together into a separate analytics environment to do analysis which can be an analytical server or a personal computer with more computing capability.

