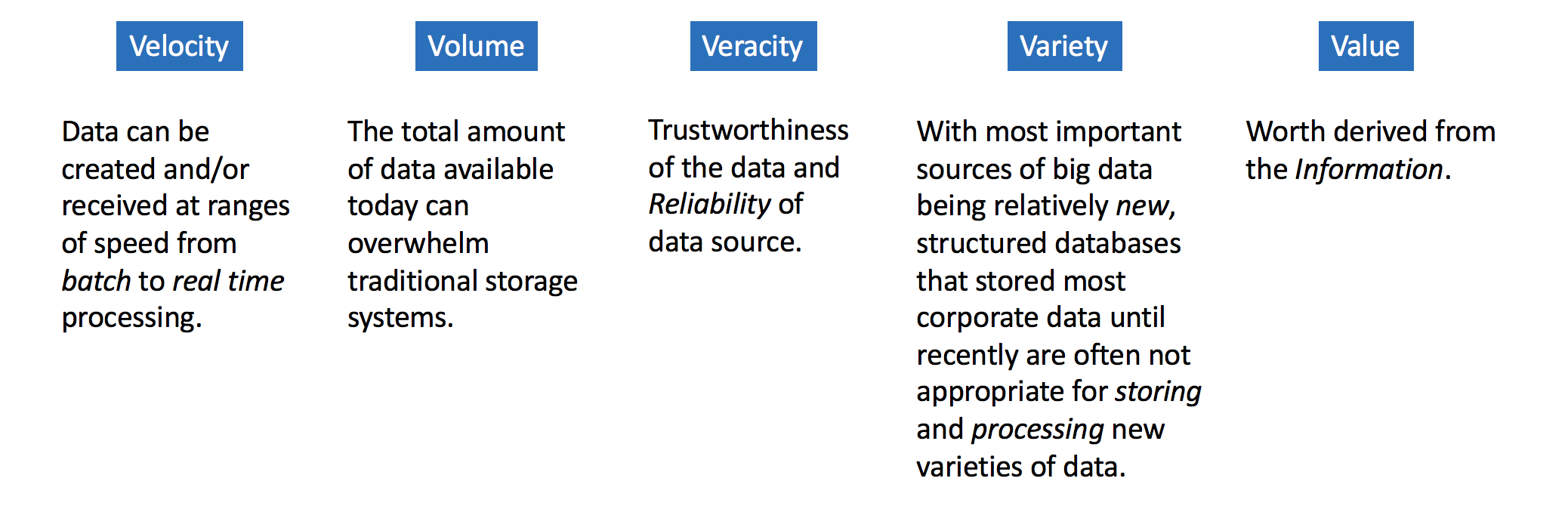
Big Data and Analytics Toolkit

Over the last 30 years, the amount of information, devices and softwares have exponentially increased. “Big Data” is the coined term that characterizes this growing data revolution. Big Data is the change agent that challenges the ways in which organization leaders have traditionally made decisions. It is important for organizations to understand evolving big data architectures that support analytics driven solutions. It is important for organizations to choose relevant big data technologies that support data intensive applications. Lastly, organizations must be prepared and confident in deploying big data analytics initiatives. This toolkit aims to highlight key concepts from the “Certificate in Big Data and Analytics” course.

**5 Vs of Big Data:**



Organizations must consider the following “Big Data Risks”:

|  |  |  |  |
| --- | --- | --- | --- |
| **Overwhelming Ecosystem:**  Choosing the right technologies, personnel and minimize the dark data. | **Privacy:**  Must navigate government regulations, legal liabilities, and internal user privacies. (i.e. GDPR) | **Costs:**  As data grows, costs grow. It is challenging to project costs of unknown phenomenon. | **Security:**  Securing data is very challenging. Must consider securing infrastructure, staff integrity and disaster recovery. (i.e. Equifax data breach) |

When organizations consider storing massive amounts of data, there are a few most common options:

|  |  |  |  |
| --- | --- | --- | --- |
| **Object Storage:**  No directory hierarchy, cheap, good for archival data | **File System Storage:**  Hierarchy folder structure, most efficient for smaller file operations | **Database Storage:**  Heavy metadata, managed, high costs, best for connected data | **Streaming & Queue Storage:**  Ordered storage and delivery, data buffer, message routing |

Most organizations with Big Data Analytics teams seek to employ a combination of the following professional roles:

* Data Engineer
  + Manage data storage, security and platform
* Data Scientist
  + Explore data and modeler
* Data Analyst
  + Generate reports and visualizations

As data volumes grow, hardware requirements must change and scale with the data. Traditional processing environments deployed **symmetric multi-processing** systems. Scaling SMP systems required purchasing bigger servers to replace current servers (scale up). Modern architectures deploy **massively parallel processing** systems which allows for adding servers to scale with the data (scale-out).

Hadoop is one of the most popular developments associated with the open source revolution. Hadoop consists of 2 main components:

|  |  |
| --- | --- |
| **HDFS**  Distributed, replicated storage which can leverage commodity hardware. | **MapReduce**  Distributed processing paradigm based on open source frameworks. |

In the era of big data, traditional SQL and NoSQL Databases are still popular. For instance:

* RDBMS = Relational Database Management Systems
  + SQL support, fixed schema, supports transactions
  + Ex. Hive, PostgreSQL, MySQL, MSFT SQL Server
* NoSQL DBs = Not Only SQL Databases
  + Simple SQL-ish language, No fixed schema, no transaction support
  + Ex. HBASE, Cassandra

Streaming information is growing in popularity. Kafka is one of the most popular open source streaming technologies. Kafka is a message broker based on distributed cluster system. Many popular cloud vendors offer a proprietary message broker application.

Organizations must implement an appropriate and cost effective data storage strategy. The most common data storage strategies are data warehouses and data lakes.

|  |  |  |
| --- | --- | --- |
| **Data Warehouse**  Highly organized and structured. Subject specific data store | **Data Mart**  Small scale DW. Subject/department specific data store | **Data Lake**  No organization required. Raw data content. |

Cloud platforms are increasing in popularity. The 3 Major Cloud Platforms:

* Amazon Web Services
* Microsoft Azure
* Google Cloud

**Apache Spark** is an open source, distributed in-memory based processing engine. Spark was developed to address the inefficiencies of Hadoop MapReduce. Apache Spark is the most popular Apache project in the world, and has a great developer community. Databricks is a PaaS-like service that leverages Spark as it’s processing engine.

Today, organizations can leverage big data technologies to power artificial intelligence and analytics applications.

|  |  |
| --- | --- |
| **Artificial Intelligence**  Any technique that enables a computer to mimic human behavior. Deep learning is a common technique that extends the neural network framework. | **AI Ethics**  Ethical approaches to AI are essential. Avoid discrimination, racial and social biases to ensure consumer’s confidence and good faith. |