**Introduction to Big Data**

* **IBM gives a definition for BIG DATE that any data that has 3 V's that can be called as BIG DATA. Those 3v's are Volume, Variety, and Velocity.**
* **Usually in traditional data systems we previously deal with only structured data when it comes to Big Data it is more of Structured, Semi Structured and Unstructured Data as well.**
* **Volume: When it comes to volume the Data is very huge now a days that traditional system couldn’t be able to handle. In general, around 2.5 quintillion bytes of data is being generated every day.**
* **Variety: the data could be structured, semi structured and unstructured.**

**Structured (RDBMS Data bases (Oracle and MySQL)**

**Semi structured (CSV, XML, JSON)**

**Unstructured (Audio, Video, Image and Log Files)**

* **Velocity: The Systems should be able to tackle huge traffic that comes in to.**

**The Whole goal of BIG DATA is to process huge amount of data which traditional systems are not capable of processing which can be difficult by using traditional data systems.**

**Before processing, there is another huddle called storage, as to process huge amount of data we first need to store it! Traditional systems are not capable of storing large or massive amount of data.**

**STORE 🡺 PROCESS 🡺 SCALE**

* **Store (store massive amount of data)** 🡺 **PROCESS (process it in a timely manner)**🡺 **SCALE (Scale easily as data grows)**

**These are the three things required to design a good big data system.**

* **When it comes to scalability, there are two ways to design system 1. Monolithic and 2. Distributed.**
  + - * **1.Monoloithic** 🡺 **One powerful system with lot of resources.**
      * **2. Distributed**🡺 **Many smaller systems come together.**
* **In monolithic as it is a single powerful server, it is hard to add resources after a certain limit. Resources could be RAM, HARD DISK, CPU.**
* **Monolithic is not scalable that means if monolithic system becomes 2X resources != 2X Performance. So that’s why monolithic system cannot be scalable after certain limit.**
* **On the other hand, imagine we have 6 nodes which is a distributed system, and it is scalable so in distributed systems, 2X resources = 2X speed. That is why all the big data systems are distributed system because of this scalability after a certain point. Which is no in the case of Monolithic systems.**
* **Monolithic architecture is based on Vertical scaling (Not True Scaling), Distributed scaling is based on Horizontal scaling.**
* **So in summary here we are trying to solve the issues we face in case of Big data, as initially we need to deal with store after that processing and then scalability (for this we use distributed systems).**