

Data modelling is on discussion in every table. We are yet to provide correct way to identify the root analysis and provide solution according to it. I , being a leaner , hover around several tutor and youtube training videos and try to listen to the master minds about their thought process. Hence came along with **“BIG DATA and BREWS”** where they provide a simplistic view about the same.

**“The Data”** been an asset to any organization, it is of prior importance to understand how to manage and mitigate the risks at the same time.

With no doubts, the storage and CPU is becoming cheaper especially, with the cloud providers like Amazon, EMC and Microsoft encouraging client to shift from on-premise to cloud server. In our market terms, we call them hyperscaler disrupting all the traditional way of looking into the things. The next question comes, what we will do now. Is it only defined to data or application as well. How about the Web API world developing at parallel. For example : I-Store of I-phone trying million of apps every now and then or say twitter providing Web application easily registered and encourages small business.

I don’t argue to the fact, that CRM and ERP world with remain same as it before. Mostly important if you understand the fact that these applications run on static schema or schema bound, where data is the subset of the application. Thus RDBMS like Oracle or Mysql plays an important role with no replacement is near future.

But on the same note, there is another evolving at parallel with other way of looking into the things. The Data been an asset, to any organization, (where application come and goes), this world, can be divided into 3 layers of Data Abstraction.

1. **DATA PERSISTENCE –**

This layer of abstraction is all about understanding understanding the data sources. In accordance to all the data world master minds who I listen to ,( Thanks to google), the data sources can be divided in 3 ways.

1. **Volume** - Clients with huge volume of data and is in dilemma to store in correct and best possible
2. **Variety** - Data being structured, Unstructured, Quasi or Semi Structure , documented , what is the best possible way to handle/manage it.
3. **Velocity** - How fast the data is appended and how to bring insights to the same.

The sources can be anything including the remote sensor data, web log data, weather report data or even document store data. It is as important to organization as organized/structured data especially taking into consideration the security and vulnerability.

Thus , it is of prior importance of all the data master minds who we follow , to bring them is single layer and not being a subset to application. In current world we call it as Data Lake or Data River.

The ideal way of looking into the things was always as Extract , Transform (Say Development/Staging Data) and Load into the system. These master minds are dwelling more into the fact that with more introduction to cloud service and disrupting the current architecture, we can go with **extract, load the data and then transform at the end user level.**

1. **DATA INFRASTRUCTURE**

While reading and briefing more about the second layer of abstraction, this layer belongs to many renown companies whose prior enjoy is their processing engine. Some with big names like Informatica, Alteryx , Talent focusing more on this layer.

This layer can be sub divided into categories mentioned below :

* Data Ingestion
* Data Cleansing
* Transforming and processing quality data
* Security with single pivotal command center
* Bringing Application to the Data.

We classify the data into 3 categories based on the sources from where it is coming. These are :

1. Static Data
2. Streaming Data
3. Real Time Data

Big Data World belong to Hadoop/MR is the past with batch processing being the headache when we consider the volume and variety of data. The Hadoop/MR being the core component of the ecosystem in terms of processing engine, the world seen other auxiliary components as Hive/Pig/Mahout based on requirement/competency skill and the choices of the company like Facebook, yahoo etc.

But then the intelligent data world says, as time flows , so does the data flows. Thus we have Velocity/Streaming Data coming into the picture. To bring everything into single stack, we introduce SPARK or in memory data processing which goes well with Hadoop/HDFS file system as well.

Once we can identify, organize and cleaning the data , we have resilient distributed data sets in SPARK concept. Based on our requirement it goes well with dataframe/sql, streaming data or graph data.

Some of the renown companies as already mentioned like Informatica/Alteryx, thus introduced SPARK along with their processing engine to stay competitive in the market.

1. **DATA INTELLIGENCE**

This layer is the 3rd layer of data abstraction, which was earlier impossible but growing fastly talks all about Data Visualization and Machine Learning Algorithms.

Major players like Amazon, OTTO , google use search and recommendation engine to find the correct mapped data.

Also , companies like Tableau and Power BI solely focus on visualization of data and creating of dashboard for Sales /Marketing and Business minds.

Where we focus on Data most of the time, API/Application user world is evolving at parallel and trying out new application is becoming the new trend. All the mega minds / master minds are talking of one thing in common as Data being an independent layer and bringing application to the data in the form of swamp as of when required.

Ted Talk/Youtube/Tutor Videos are good data streaming sources helped me to idealize the concept. I entirely admire them all.