**Task 1**

Consider the usage examples of Big Data for a car company. Assume that company manufactures five models of cars, and each model is available in five colors and five shades. The company collects inputs from customers and sales centers, and inputs of component malfunctions from service centers for different models. The company also uses social media inputs.

Explain 3Vs characteristics in this company's data.

**Ans:**

The 3Vs of big data is,

**1.** Volume

**2.** Velocity

**3.** Variety

**Volume:**

Here, I have taken the example of a car company. Nowadays car has been improved and technological advanced like,

* Update maps automatically.
* Choose the best route.
* Adjust internal climate based on weather conditions.
* Analyze engine efficiency.
* Update a music playlist and many more.

This type of activity will produce lots of data and amount of data is even so big it’s pretty much impossible to picture.

**Velocity:**

I have talked about the volume of data above but this data is also produced so fast. Because as Google 1.4 billion cars are on the road and all that cars produced lots of data in every day.

Here, I have only talked about data produced by only car but if I talk about social media, customer reviews and sales centers than more than 10 million cars has been sold per year. So, the data produced by every area is very huge in number.

**Variety:**

The data produced by this car company is in different form. It’s not in traditional row and column format because data has pictures, videos, number data, description, customer reviews, sales graph, music etc.

**Task 2**

Study the Drivers of Big Data and Summarize your Understanding.

**Ans:**

**1. Data driven initiatives:**

They are primarily categorized into 3 broad areas:

**a. Data Driven Innovation:** I particularly like the innovation aspect with being data driven. Imagine being able to learn from your customer first what they need and having the ability to drive innovation through those uber targeted data indicators.

**b. Data Driven Decision Making:** Data driven decision-making is the inherent ability of analytics to sieve through globs of data and identify the best path forward. Whether in terms of finding the best route to validating the current route and estimating the success/failure in current strategy. It takes decision making away from gut and focus on data backed reasoning for higher chances of success.

**c. Data Driven Discovery:** Your data know a whole lot about you than you image. Having a discovery mechanism will help you understand hidden insights that were not visible through traditional means.

**2. Data Science as a competitive advantage:**

I had the fortune of interacting with couple of mid size company’s executives from commodity businesses. There had been a consistent outcry on having to build big data as a capability to add to their competitive advantage. With a proper data driven framework, businesses could build sustainable capabilities and further leverage these capabilities as a competitive edge. If businesses were able to master big data driven capabilities, businesses could use these capabilities to establish secondary source of revenues by selling it to other businesses.

**3. Sustained processes:**

Data driven approach creates sustainable processes, which gives a huge endorsement to big data analytics strategy as a go for enterprise adoption. Randomness kills businesses and adds scary risks, while data driven strategy reduces the risk by bringing statistical models, which are measurable.

**4. Cost advantages of commodity hardware & open source software:**

Cost advantage is music to CXO’s ears. How about the savings your IT will enjoy from moving things to commodity hardware and leverage more open source platforms for cost effective ways to achieve enterprise level computations and beyond. No more overpaying of premium hardware when similar or better analytical processing could be done using commodity and open source systems.

**5. Quick turnaround and less bench times:**

Have you dealt with IT folks in your company? Mo and mo people, complex processes and communication charter gives you hard time connecting with someone who could get the task done. Things take forever long and cost fortunes with substandard quality. A good bigdata and analytics strategy could reduce the proof of concept time smoothly and substantially. It reduces the burden on IT and gets more high quality, fast and cost effective solutions baked. So, you will waste less time waiting for analysis / insights and more time digging through mo and mo data, and use it for better insights and analyses which was never heard of before.

**6. Automation to backfill redundant/mundane tasks:**

How about doing something to the 80% of time that is wasted in data cleaning and preprocessing. There is great deal of automation that could be take part and sky rocket enterprise efficiency. Less manual time spent on data prep and more time is spent on doing analysis that would have substantial ROI compared to mundane data preps and monotonous tasks.

**7. Optimize workforce to leverage high talent cost:**

This is an interesting area that I am keeping a close eye on. Businesses already have right talent pools that would solve some pieces of the big data puzzle on data science. Businesses have BI, Modelers and IT people working in harmony in some shape or form. So, a good big data & analytics strategy ensures current workforce is leveraged to it’s core in handling enterprise big data and also ensures right number of data scientists are involved with clearer sight to their contribution and their ROI.

**8. Data continues to grow exponentially:**

Whether you like it or not, data is increasing. One key technological push is the increasing data and the threat of not being able to use this exploding enterprise data for insights. Having a good strategy puts a pacifier to growing unutilized data concerns.

**9. Data is everywhere and in many formats:**

Besides being able to sieve through data in huge volumes, having a stream of disparate data also poses its threats. Text, voice, video, logs and other emerging formats make it harder to gain insights using traditional tools. So, businesses need to drive their big data toolkit to prep for this exploding data type that is entering corporate data DNA.

**10. Alternate, Multiple Synchronous & Asynchronous data streams:**

Data coming through multiple silos in real time, creating problem in keeping up with this data in existing data systems. These multiple streams put pressure on businesses to have an effective strategy on handling these sources. With tools out there to handle such situations, it has become important to acquire such capabilities before the competition does.

**11. Low barrier to entry:**

As with any business, low barrier to entry poses one great leverage for businesses to try different technologies and come up with the best strategy. Easy frameworks & paradigms have made available lots of tools, which are relatively easier to deploy. These tools could deliver a phenomenal computing horsepower.

**12. Traditional solutions failing to catch up with new market conditions:**

Big data has given rise to exploding volume, velocity and variety of data. These 3Vs are difficult to handle and demand cutting edge technologies. New requirements have emerged from changing market dynamics that could not be addressed by old tools, but demands new big data tools. Hence, a big data and analytics strategy to embrace these tools before business goes obsolete.

**Task 3**

Study the Data Formats given below and summarize your understanding.

**Ans:**

**1. CSV/TSV:**

Full form of CSV is comma separated values and TSV is tab separated values. In CSV data is stored in row format and values are separated by comma. In TSV data is stored in row format and values are separated by tab. Extension of CSV file is .csv and TSV file is .tsv.

**2. JSON**

JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others. These properties make JSON an ideal data-interchange language.

**3. Parquet**

Parquet is an open source file format available to any project in the Hadoop ecosystem. Apache Parquet is designed for efficient as well as performant flat columnar storage format of data compared to row based files like CSV or TSV files.

**4. Avro**

Avro is an open source project that provides data serialization and data exchange services for Apache Hadoop. Avro stores the data definition in JSON format making it easy to read and interpret; the data itself is stored in binary format making it compact and efficient.

**5. Key-Value Pair**

A key-value pair consists of related elements: A key, which is a constant that defines the data set (e.g., gender, colour, price) and a value, which is a variable that belongs to the set (e.g., male/female, green, 100). Destination Builder sends data formatted as key-value pairs.

**Standard and Serial Key-Value Pairs**

Destinations accept key-value data in standard or serialized format.

**a. Standard key-value pairs:** Formats destination data into separate key-value pairs. Each key is stated explicitly, even when used again to define a different value.

**b. Serialized key-value pairs:** Condenses multiple values into a single key-value pair. In a serialized key-value pair, a special indicator separates the values within the key-value set.

Both standard and serialized key-values can contain single or multiple values