**Task\_1**

What are tools for big data?!

1. **Hadoop:**

Pros :

Very useful for research and development purposes.

Offers easy data access.

Extremely scalable.

Cons:

Data redundancy can often cause disk space problems.

For improved efficiency, I/O operations should have been optimized.

Pricing : free

1. **Xplenty:**

Pros:

It is a cloud network that is elastic and scalable.

It offers a customized and flexible API component.

By using the rich expression language of Xplenty, you can incorporate complex data preparation functions.

Cons:

There is no option for monthly subscription.

Pricing: It has a price model focused on subscriptions and can be tried for free for 7 days.

1. **CDH( Cloudera Distribution for Hadoop):**

Pros:

Complete and accurate distribution.

The Hadoop cluster is very well managed by the Cloudera Manager.

Simple to deploy.

The administration is less complicated.

High security and administration

Cons:

Several suggested installation methods are confusing.

Pricing: Cloudera edition of CDH is a free Big Data Analytics tool.

1. **Cassandra :**

Pros:

There is no single failure point.

It manages huge data really quick.

It has log-structured storage and linear scalability.

Cons:

Extra troubleshooting and maintenance work is required.

1. **Datawrapper:**

Pros:

Operates exceptionally well on any type of device – smartphone, laptop, or tablet.

Rapid and interactive responses.

Excellent export and customization options.

Cons:

Has limited options for color palettes.

Pricing: it offers free servers**.**

1. **MONGODB :**

Pros :

Supports various platforms and technologies.

No install and maintenance hiccups.

Robust and cost-effective.

Cons:

it has a limited analytics resource.

1. **Tableau:**

Pros:

No-code queries and visualization

Easy setup

Real-time collaboration

Straightforward integrations

Cons:

More expensive than some tools

Customer support frustrations

Pricing: For desktop, servers, and online, Tableau offers various editions. Its price begins at $35 a month. A free trial is available in any edition.

1. **Apache spark:**

Pros:

Open-source

High-level operators

More flexibility and versatility than Hadoop

Supports real-time and batch processing, plus in-memory calculations

Cons:

Advanced training required

Documentation not always helpful

Extra security measures required

1. **Microsoft Azure**
2. **RapidMiner :**

offers the most powerful and intuitive graphical user interface for the design of the analysis process.

**Task\_2**

What are the 7 v’s of Big Data?!

* Volume : how much data we have,  measured in Gigabytes is now measured in Zettabytes (ZB) or even Yottabytes (YB).
* Velocity : the speed in which data is process and becomes accessible.
* Variety :  It can be unstructured and it can include so many different types of data from XML to video to SMS. Organizing the data in a meaningful way is no simple task, especially when the data itself changes rapidly.
* Variability:
* Veracity:  making sure the data is accurate.
* Visualization: Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets and reports chock-full of numbers and formulas.
* Value : After addressing volume, velocity, variety, variability, veracity, and visualization – which takes a lot of time, effort and resources – you want to be sure your organization is getting value from the data.

**Task\_3**

P- VALLUE & CONFIDENCE INTERVAL

|  |  |
| --- | --- |
| P-VALUE | Confidence interval |
| calculated to assess whether trial results are likely to have  occurred simply through chance  provide a cut-off beyond which we assert that the  findings are ‘statistically significant’ (by convention, this is p<0.05). | calculated for a measure of treatment effect  shows the range within which the true treatment effect is likely to lie  preferable to p-values, as they tell us the range  of possible effect sizes compatible with the data. |