Explorative Study of Data Science Model and Its Tools Set Comparison

**Jibran Rasheed Khan1,2, Idrees Riaz2, Muhammad Waqar2**

**1,2Lecturer at Department of Computer Science – UBIT, and Research Officer at HEJ, ICCBS, University of Karachi**

**2Department of Computer Science Information & Technology – CSIT, NED University of Engineering and Technology.**

**Abstract**

Data Science (DS) is a scientific technique to unfold the hidden mystery of data and extract the valuable insight that can boom the business needs, beware of what probably be happened in near future and most importantly make human proactive to act than react on events. Nowadays, data is the key element that use to classify and aids in to foretell each and every aspect of human; behavior, nature, standards, lifestyle, passion, desire, business needs and lot more uncountable. But, data is not so simple nor resides in any specified form. This usually be available in any of 4Vs; Volume, Velocity, Verity and Veracity. Due to complex nature of data, many organization, scientist and practitioners suffer trouble to incorporate them in a suitable platform, desirable tools, libraries and other supportive algorithms or calculations. Also, data science itself not a separate field of study, but it’s a composition of several incorporating fields – Information Technology, Statistics and Science. The core propose of this explorative study is focused on the learning DS generic model and comparison of its tools set in the market. In this work, we will explore each aspect and compositor of data Science, drill down to identify individual field hierarchy, tools, and platform available. Furthermore, it covers the concise comparison study renewed and most preferable element in each filed. The study, will beneficial in identifying actual requirements and composite tool that will ease the practitioners to learn more, building better and flexible system and more compatible integrating features.

**Introduction**

**Data Analysis Tools**

Although, the data science is not much aged, but it captured attention of everyone across the world. This is the reason lots of companies take part in building handy tools to do the complicated tasks comfortably by offering wide range of tools. It is impossible to list every tool, we focus on some most prominent and high rated tools in the following section.

1. **Microsoft HDInsight**

Azure HDInsight is a Spark and Hadoop service in the cloud. It provides big data cloud offerings in two categories, Standard and Premium. It provides an enterprise-scale cluster for the organization to run their big data workloads. Azure HDInsight is an easy, cost-effective, enterprise-grade service for open source analytics that enables customers to easily run popular Apache open source frameworks including Apache Hadoop, Spark, Kafka, and others. The service is available in 27 public regions and Azure Government Clouds in the US and Germany.

Azure HDInsight powers mission critical applications in a wide variety of sectors and enables a wide range of use cases including ETL, streaming, and interactive querying.

# Features

* Reliable analytics with an industry-leading SLA
* It offers enterprise-grade security and monitoring
* Protect data assets and extend on-premises security and governance controls to the cloud
* High-productivity platform for developers and scientists
* Integration with leading productivity applications
* Deploy Hadoop in the cloud without purchasing new hardware or paying other up-front costs