*Spark*

**A C A D G I L D Page 1**

Session 1: Introduction to Big Data

Assignment 1*Spark*

**A C A D G I L D Page 2**

*Session 1: Introduction to Big Data*

*Assignment 1*

**Table of Contents**

1. Introduction ........................................................................................................................................ 2

2. Objective ............................................................................................................................................. 2

3. Prerequisites ....................................................................................................................................... 2

4. Associated Data Files ............................................................................................................................ 2

5. Problem Statement ............................................................................................................................... 2

6. Approximate Time to Complete Task ................................................................................................... 3

**1. Introduction**

In this assignment you need to select the correct choice for the given questions

**2. Objective**

This assignment will help you to understand Big Data basics.

**3. Prerequisites**

None

**4. Associated Data Files**

N/A

**5. Problem Statement**

1. Hadoop was developed by

a) Doug Cutting

b) Lars George

c) Tom White

d) Eric Sammer *Spark*

*Ans:a)Doug Cutting*

**A C A D G I L D Page 3**

2. What mechanisms does Hadoop use to make NameNode resilient to failure?

a) take a backup of a filesystem metadata to a local disk and a remote NFS mount

b) use a machine with at least 12 CPUs

c) store the filesystem metadata in a cloud

d) using expensive and reliable hardware

Ans : a) take a backup of a filesystem metadata to a local disk and a remote NFS mount

3. What kind of servers are used for creating a Hadoop cluster?

a) server grade machines

b) commodity hardware

c) only supercomputers

d) none of the above

Ans : b) commodity hardware

4. As compared to RDBMS, Hadoop has

a) has higher data integrity

b) does ACID transactions

c) is suitable for read and write many times

d) works better on unstructured and semi-structured data

Ans: d) works better on unstructured and semi-structured data

5. When do you choose big Data tools for processing?

a) if you have more no of spare machines

b) when the input data volume is around 10 GB to 50 GB

c) Hadoop is a replacement tool for Java or DotNet applications

d) when you encounter processing challenges in terms of Volume, Velocity and Variety

Ans : d) when you encounter processing challenges in terms of Volume, Velocity and Variety

**6. Approximate Time to Complete Task**

10 min

*Spark*

**A C A D G I L D Page 1**

Session 1: Introduction to Big Data

Assignment 2*Spark*

**A C A D G I L D Page 2**

*Session 1: Introduction to Big Data*

*Assignment 2*

**Table of Contents**

1. Introduction ........................................................................................................................................ 3

2. Objective ............................................................................................................................................. 3

3. Prerequisites ....................................................................................................................................... 3

4. Associated Data Files ............................................................................................................................ 3

5. Problem Statement ............................................................................................................................... 3

6. Approximate Time to Complete Task ................................................................................................... 3 *Spark*

**A C A D G I L D Page 3**

**1. Introduction**

In this assignment you need to answer the given questions.

**2. Objective**

This assignment will help you to understand Big Data basics.

**3. Prerequisites**

None

**4. Associated Data Files**

N/A

**5. Problem Statement**

1. What are the major sources of Big Data?

Ans :Log Data,SensorData,Image Processing data,Telecome data,Airline data,Social Media.

2. What is the criteria to decide whether the data is Big Data?

Ans : Volume,Variety,Velocity(V3 rule)

3. Give three examples for unstructured data?

Ans : Image/Audio,Twitter data,Analog data,GPS traking.

4. What is streaming data?

Ans : Streaming Data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously, and in small sizes (order of Kilobytes)

Uses :

Complex Event Processing Network Monitoring

Risk Mannagement

Streaming Data are :E-commerce,Froud Detection,Smart order routing,Algorithm Trading.

5. What is the difference between Batch and Streaming data?

Straming Data: Intermediate processing of data after the event occours.

Real time

Performance and Volumne challenges

Usecase :Operation/Analytics

**6. Approximate Time to Complete Task**

10 min

*Spark*

**A C A D G I L D Page 1**

Session 1: Introduction to Big Data

Assignment 3*Spark*

**A C A D G I L D Page 2**

*Session 1: Introduction to Big Data*

*Assignment 3*

**Table of Contents**

1. Introduction ........................................................................................................................................ 2

2. Objective ............................................................................................................................................. 2

3. Prerequisites ....................................................................................................................................... 2

4. Associated Data Files ............................................................................................................................ 2

5. Problem Statement ............................................................................................................................... 3

6. Approximate Time to Complete Task ................................................................................................... 3

**1. Introduction**

In this assignment you need to answer the given questions.

**2. Objective**

This assignment will help you to understand Big Data basics.

**3. Prerequisites**

None

**4. Associated Data Files**

N/A *Spark*

**A C A D G I L D Page 3**

**5. Problem Statement**

1. Give a brief description of the Hadoop eco system?

The Hadoop platform consists of two key services: a reliable, distributed file system called Hadoop Distributed File System (HDFS) and the high-performance parallel data processing engine called Hadoop MapReduce, described in MapReduce below. Hadoop was created by Doug Cutting and named after his son’s toy elephant. Vendors that provide Hadoop-based platforms include Cloudera, Hortonworks, MapR, Greenplum, IBM, and Amazon.

Hadoop Ecosystem : pig,hive,spark,flume,Hbase,mahout etc.

2. What is meant by distributed computing and what are the different types of scaling available?

**Distributed computing** is a field of **computer** science that studies **distributed** systems. A **distributed** system is a model in which components located on networked **computers** communicate and coordinate their actions by passing messages. The components interact with each other in order to achieve a common goal.

Scaling is one of the issue with convvesional system.

Scale up:Adding resources into the single node system.

Scale out:Adding more nodes to a system.

More challenging for fault tolerance and Software Development.

3. What is meant by commodity hardware in real time and what are its advantages with reference to enterprise?

Commodity hardware, in an IT context, is a device or device component that is relatively inexpensive, widely available and more or less interchangeable with other hardware of its type.

its advantages with reference to enterprise is

1)scalable

2)cost-effective

3)Flexible

4)Fast

5)Resilient to failure

4. What is meant by real time and how data is collected in the real time and what are the options for analyzing streaming data?

Real Time: Immediate processing of data after the events occurs.Performance and Volume challening.

The data is collected into the real time system by streaming.

Apchae spark,Strom these are the Stream processing tools.

5. What is the difference between HDFS blocks and input splits?

**Block** and **split** size can be changed in properties. Map reads data from **Block** through **splits** i.e. **split** act as a broker **between Block** and Mapper. ... **Input splits** are a logical division of your records whereas **HDFS blocks** are a physical division of the **input** data

**6. Approximate Time to Complete Task**

10 min