**20IT701**

**Hall Ticket Number:**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |

**BIG DATA ANALYTICS**

**IV/IV B.Tech (Regular/Supplementary) DEGREE EXAMINATION**

**December, 2025** **IT**

**Seventh Semester**

**Time:** Three Hours **Maximum:**70 Marks

***Answer question 1 compulsorily.*** **(14X1 = 14Marks)**

***Answer one question from each unit.*** **(4X14=56Marks)**

| | | | CO | BL | M |
|---|---|---|---|---|---|
| 1 | a) | What is Big Data? | CO1 | L1 | 1M |
| | b) | Give one characteristic of Big Data. | CO1 | L1 | 1M |
| | c) | Name two sources of Big Data | CO1 | L1 | 1M |
| | d) | Define Big Data Analytics. | CO1 | L1 | 1M |
| | e) | Expand YARN. | CO2 | L1 | 1M |
| | f) | What is the main role of YARN in Hadoop? | CO2 | L2 | 1M |
| | g) | Name two main components of YARN. | CO2 | L1 | 1M |
| | h) | Mention one difference between MapReduce v1 and YARN. | CO2 | L2 | 1M |
| | i) | What is Pig? | CO3 | L1 | 1M |
| | j) | What is Pig Latin? | CO3 | L1 | 1M |
| | k) | What are the two execution modes of Pig? | CO3 | L1 | 1M |
| | l) | What is Hive QL? | CO3 | L1 | 1M |
| | m) | How does Sqoop connect to databases? | CO4 | L1 | 1M |
| | n) | What is Sqoop used for? | CO4 | L1 | 1M |
| | | **Unit-I** | | | |
| 2 | a) | Differentiate between traditional databases and Big Data systems. | CO1 | L4 | 7M |
| | b) | Discuss about three applications of Big Data Analytics. | CO1 | L2 | 7M |
| | | **(OR)** | | | |
| 3 | a) | Describe the role of the Name Node and Data Node in HDFS. | CO1 | L3 | 7M |
| | b) | How does HDFS achieve fault tolerance? | CO1 | L2 | 7M |
| | | **Unit-II** | | | |
| 4 | a) | What is scheduling in YARN? discuss its importance. | CO2 | L2 | 7M |
| | b) | Explore the anatomy of a MapReduce job run. | CO2 | L3 | 7M |
| | | **(OR)** | | | |
| 5 | a) | Explain the Shuffle and Sort phase in MapReduce with example | CO2 | L3 | 7M |
| | b) | Describe the role of sorting in MapReduce and present a use case. | CO2 | L3 | 7M |
| | | **Unit-III** | | | |
| 6 | a) | Compare Hive and RDBMS in detail. | CO3 | L4 | 7M |
| | b) | Write about Managed and External Tables with examples. | CO3 | L3 | 7M |
| | | **(OR)** | | | |
| 7 | a) | Explain Hive Query Flow give an example. | CO3 | L3 | 7M |
| | b) | discuss Hive UDFs and their significance using an example | CO3 | L3 | 7M |
| | | **Unit-IV** | | | |
| 8 | a) | Explain the complete process of importing data using Sqoop. | CO4 | L3 | 7M |
| | b) | Write about three transformations supported by Spark give examples. | CO4 | L3 | 7M |
| | | **(OR)** | | | |
| 9 | a) | Describe the Spark Architecture with a neat diagram and explain the roles of the Driver and Executor. | CO4 | L4 | 7M |
| | b) | Explain the process of creating and transforming RDDs in Spark with examples. | CO4 | L3 | 7M |

## SCHEME OF EVALUATION

**Section–A (1 × 14 = 14 Marks)**

Each question carries **1 mark**.
Award full marks if keywords are present; give 0.5 mark for partially correct statements where applicable.

| Q.No | Expected Answer (Key Points) | Marks |
|---|---|---|
| 1(a) | Definition of Big Data – large, complex datasets that traditional systems can't handle | 1M |
| 1(b) | Any one characteristic: Volume / Velocity / Variety / Veracity / Value | 1M |
| 1(c) | Any two sources: Social media, sensors, logs, IoT, transactions, machines | 1M |
| 1(d) | Definition of Big Data Analytics – analysis of Big Data to discover patterns/insights | 1M |
| 1(e) | YARN – Yet Another Resource Negotiator | 1M |
| 1(f) | Role of YARN – resource management + job scheduling | 1M |
| 1(g) | Two components: Resource Manager, Node Manager (also Application Master, Container) | 1M |
| 1(h) | One difference: In v1, JobTracker does both RM + scheduling; YARN separates RM/AM | 1M |
| 1(i) | Pig – high-level data flow scripting platform on Hadoop | 1M |
| 1(j) | Pig Latin – language used in Apache Pig | 1M |
| 1(k) | Modes: Local mode, MapReduce mode | 1M |
| 1(l) | HiveQL – query language for querying data in Hive | 1M |
| 1(m) | Sqoop connects using JDBC | 1M |
| 1(n) | Sqoop used for transferring data between Hadoop and RDBMS (import/export) | 1M |

---

**Section–B (Unit-wise 14 Marks Each)**

Each sub-question carries **7 Marks**.
Use the following marking scheme:

**General Marking Pattern (7M questions)**

- **Definition / Concept clarity:** 1–2 marks
- **Explanation / Description:** 3–4 marks
- **Diagram / Example / Comparison / Steps:** 1–2 marks
- **Total:** 7 marks

---

**Unit–I (Q2 or Q3)**

**Q2(a) Differentiate traditional DB vs Big Data systems – 7M**

- Minimum 5–7 points of comparison (architecture, scalability, data type, processing, schema, cost, etc.)

**Q2(b) Applications of Big Data Analytics – 7M**

- Any 3 applications with explanation: healthcare, finance, marketing, IoT, e-commerce, fraud detection, etc.

---

**Q3(a) NameNode & DataNode roles – 7M**

- NameNode: metadata, namespace mgmt, file system tree, block mapping
- DataNode: stores blocks, serves read/write requests

**Q3(b) Fault tolerance in HDFS – 7M**

- Replication, heartbeat, block reports, rebalancing, rack awareness

---

**Unit–II (Q4 or Q5)**

**Q4(a) YARN scheduling + importance – 7M**

- FIFO/Fair/Capacity scheduling, how it allocates resources, importance in cluster utilization

**Q4(b) Anatomy of MapReduce job run – 7M**

- Steps: Input split → Map → Shuffle → Sort → Reduce → Output
- Mention JobTracker + TaskTracker roles or RM/AM roles in MRv2

---

**Q5(a) Shuffle & Sort phase with example – 7M**

- Detailed explanation of shuffle, sort, partition, merge; example word count

**Q5(b) Role of sorting in MapReduce + use case – 7M**

- Sorting groups keys for reducers; required for deterministic output
- Use case: log analysis, ranking, frequency counting

---

**Unit–III (Q6 or Q7)**

**Q6(a) Compare Hive vs RDBMS – 7M**

- At least 6 comparison points: schema, execution, latency, storage, language, OLAP/OLTP nature, indexing

**Q6(b) Managed vs External Tables – 7M**

- Definitions, differences, examples, storage behavior when dropping tables

---

**Q7(a) Hive Query Flow with example – 7M**

- Steps: parsing → compilation → optimization → execution
- Include simple SELECT example

**Q7(b) Hive UDFs + significance – 7M**

- Types: built-in, user-defined
- Advantages & sample UDF example

---

**Unit–IV (Q8 or Q9)**

**Q8(a) Sqoop import process – 7M**

- Steps: connectivity → JDBC → mapping → parallel import → HDFS/Hive load
- Syntax example

**Q8(b) Spark transformations (any 3) – 7M**

- map(), filter(), flatMap(), distinct(), union(), etc. with examples

---

**Q9(a) Spark Architecture + diagram – 7M**

- Driver, Executors, Cluster Manager, DAG, tasks
- Neat block diagram (text description acceptable)

**Q9(b) Creating + transforming RDDs – 7M**

- Creation: parallelize(), read from HDFS
- Transformations: map, filter, reduceByKey
- With examples

**DETAILED SCHEME OF VALUATION**

**Course Outcome (CO)-wise & Bloom's Level (BL)-wise Evaluation**

---

**SECTION – A (1 × 14 = 14 Marks)**
**Each question carries 1 mark.**
Award **1 mark for correct definition / keyword / correct pair of terms**.
Award **0.5 mark** for partially correct answers.

---

**1(a) What is Big Data? (CO1, L1 – 1M)**
Expected Key Points:
- Extremely large datasets
- Cannot be processed using traditional DBMS
- High volume, velocity, variety

**Marking:**
- Correct definition → **1M**
- Partial definition → **0.5M**

---

**1(b) One characteristic of Big Data (CO1, L1 – 1M)**
Any **one**: Volume, Velocity, Variety, Veracity, Value.
**Marking:**
- Any valid characteristic → **1M**

---

**1(c) Name two sources of Big Data (CO1, L1 – 1M)**
Any **two**: Social media, sensors, IoT, transactions, log files, clickstreams, machines.
**Marking:**
- Two correct → **1M**
- One correct → **0.5M**

---

**1(d) Define Big Data Analytics (CO1, L1 – 1M)**
Expected Points:
- Process of analyzing Big Data
- Discover patterns, correlations, insights

**Marking:**
- Clear definition → **1M**

---

**1(e) Expand YARN (CO2, L1 – 1M)**
Expected: **Yet Another Resource Negotiator**
**Marking:**
- Correct expansion → **1M**

---

**1(f) Main role of YARN (CO2, L2 – 1M)**
Expected: Resource management + Job scheduling.
**Marking:**
- Both keywords → **1M**
- One keyword → **0.5M**

---

**1(g) Two main components of YARN (CO2, L1 – 1M)**
Expected: Resource Manager, Node Manager, Application Master, Container.
**Marking:**
- Any two → **1M**

---

**1(h) Difference between MapReduce v1 and YARN (CO2, L2 – 1M)**
Expected:
- MRv1: JobTracker does both scheduling + monitoring
- YARN: Resource Manager + Application Master (separated roles)

**Marking:**
- Correct difference → **1M**

---

**1(i) What is Pig? (CO3, L1 – 1M)**
Expected:
- A high-level data flow platform on Hadoop.

**Marking:**
- Correct → **1M**

---

**1(j) What is Pig Latin? (CO3, L1 – 1M)**
Expected:
- Scripting language of Apache Pig.

**Marking:**
- Correct → **1M**

---

**1(k) Two execution modes of Pig (CO3, L1 – 1M)**
Expected:
- Local mode
- MapReduce mode (Hadoop mode)

**Marking:**
- Both → **1M**

---

**1(l) What is HiveQL? (CO3, L1 – 1M)**
Expected:
- Query language of Hive (SQL-like).

**Marking:**
- Correct → **1M**

---

**1(m) How Sqoop connects to databases? (CO4, L1 – 1M)**
Expected:
- JDBC (Java Database Connectivity)

**Marking:**
- Correct → **1M**

---

**1(n) What is Sqoop used for? (CO4, L1 – 1M)**
Expected:
- Import/export data between Hadoop and RDBMS.

**Marking:**
- Correct → **1M**

---

**SECTION – B (4 Units × 14 marks = 56 Marks)**
Each sub-question is **7 Marks**.
Use the below rubric for **ALL 7-mark questions**:
**Rubric for 7M Questions**

| Component | Marks |
|---|---|
| Definition / Concepts | 1–2M |
| Detailed Explanation | 3–4M |
| Diagrams / Examples / Comparison / Steps | 1–2M |
| **Total** | **7M** |

---

**UNIT – I**
**Q2(a) Differentiate Traditional DB vs Big Data systems (CO1, L4 – 7M)**
Expected 5–7 comparison points:
- Data type: Structured vs structured+semi+unstructured
- Scalability: Vertical vs horizontal
- Processing: ACID vs distributed computing
- Storage: Centralized vs HDFS
- Cost: Expensive vs commodity hardware
- Schema: Fixed schema vs schema-on-read
- Processing model: OLTP vs OLAP/Batch

**Marking Scheme:**

- Any **5–7 proper differences → 7M**
- 3–4 differences → 4–5M
- <3 differences → 2–3M

---

**Q2(b) Three applications of Big Data Analytics (CO1, L2 – 7M)**

Expected:

- Healthcare
- Banking & fraud detection
- Social media analytics
- Recommendation systems
- Smart cities
- IoT systems

**Marking:**

- 3 applications × detailed explanation → **7M**
- 3 with poor explanation → 4–5M
- <3 applications → 2–3M

---

**Q3(a) Role of NameNode & DataNode (CO1, L3 – 7M)**

Expected:

**NameNode:**

- Manages metadata
- Maintains file system namespace
- Handles block mapping
- Controls access

**DataNode:**

- Stores actual blocks
- Sends block reports
- Heartbeats to NameNode

**Marking:**

- Complete explanation → **7M**
- Partial → 4–5M

---

**Q3(b) HDFS fault tolerance (CO1, L2 – 7M)**

Expected points:

- Replication
- Heartbeats
- Block reports
- Automatic failover
- Rack awareness

**Marks:**

- 4–5 points → **7M**

---

**UNIT – II**

**Q4(a) Scheduling in YARN + importance (CO2, L2 – 7M)**

Expected points:

- FIFO, Fair, Capacity Schedulers
- Resource allocation
- Job prioritization
- Cluster utilization

**Marks:**

- Detailed scheduler explanation + importance → **7M**

---

**Q4(b) Anatomy of a MapReduce job run (CO2, L3 – 7M)**

Steps:

- Input splitting
- Map phase
- Shuffle

- Sort
- Reduce
- Output write
- Role of Resource Manager & Application Master

**Marks:**
- Stepwise flow + diagram → **7M**

---

**Q5(a) Shuffle & Sort phase with example (CO2, L3 – 7M)**
Expected:
- Partitioning
- Sorting keys
- Merging
- Grouping
- Example: Word count

**Marks:**
- Complete flow + example → **7M**

---

**Q5(b) Role of sorting + use case (CO2, L3 – 7M)**
Expected:
- Ensures key grouping
- Ordered reduce inputs
- Real-time use case: Log analysis, ranking, clicks

---

**UNIT – III**
**Q6(a) Compare Hive and RDBMS (CO3, L4 – 7M)**
Expected 6–8 points:
- Schema-on-read vs schema-on-write
- OLAP vs OLTP
- Query latency differences
- Execution engine (MapReduce/Tez/Spark)
- Data storage (HDFS)

**Marks:**
- 6+ points → **7M**

---

**Q6(b) Managed & External tables with examples (CO3, L3 – 7M)**
Expected:
- Definition
- Differences
- Storage locations
- DROP behavior
- Syntax examples

**Marks:**
- Full examples + differences → **7M**

---

**Q7(a) Hive Query Flow + example (CO3, L3 – 7M)**
Steps:
- Parsing
- Compilation
- Optimization
- Physical plan
- Execution

**Marks:**
- All steps + example → **7M**

---

**Q7(b) Hive UDFs + significance + example (CO3, L3 – 7M)**
Expected:
- Built-in UDFs
- Custom UDFs

- Testing + registering
- Example

---

**UNIT – IV**

**Q8(a) Complete Sqoop import process (CO4, L3 – 7M)**

Expected:

- JDBC connection
- Mapper allocation
- Data splitting
- Import to HDFS/Hive
- Syntax example
  sqoop import --connect jdbc:mysql://... --table ...

---

**Q8(b) Three Spark transformations (CO4, L3 – 7M)**

Expected (any 3):

- map()
- filter()
- flatMap()
- distinct()
- reduceByKey()
- union()

Each with **syntax + example**.

---

**Q9(a) Spark Architecture + diagram + roles (CO4, L4 – 7M)**

Expected:

- Driver
- Executors
- Cluster Manager
- DAG scheduler
- Diagram description

---

**Q9(b) Creating & transforming RDDs (CO4, L3 – 7M)**

Expected:

- Creation: parallelize(), textFile()
- Transformations: map(), filter(), reduce()
- Examples