OST

1. Mark the key advantages of MongoDB:

☑ It is open source
☑ It is a document-oriented, multi-purpose, distributed data store
□ It is faster than HBase
□ It is closed-source
□ It is a well-optimized key-value store
□ It is faster than Zookeeper

2. Mark the correct claims about Zookeeper:

☑ Zookeeper is a highly available process coordination service
□ Zookeeper is used in Cassandra
☑ Zookeeper is part of the Hadoop ecosystem
□ Zookeeper was developed by the National Security Agency (NSA) of the USA
□ Zookeeper is a distributed storage & analysis platform

3. Mark the correct claims about HDFS:

☑ It is a distributed file system
□ It was designed for efficient atomic reads and updates
□ The term ‘shard’ is used in the context of HDFS inter-control center fragmentation.
□ It was designed as a centralized data store

4. The Hadoop ecosystem is/supports the following:

□ Hadoop consists of node types: Name, Data, JobTracker, TaskTracker.
□ MongoDB is a component of the Hadoop ecosystem.
□ The Hadoop ecosystem consists of Zookeeper, Kibana and RDBMS.
☑ Hadoop MapReduce supports large-scale data processing.
□ Hadoop substitutes processor-based computation
☑ Example components of the Hadoop ecosystem include HDFS, HBase, Zookeeper, Pig and others.

5. HBase…:

□ …has a data model which is centered around column families
☑ …supports fast record lookup
□ …is faster then plain HDFS in all use cases
☑ …supports cell versioning
□ …is not optimized for record level insertion
☑ …supports row-level atomic updates

6. Logstash is…

☑ …about 10 years old, ie first developed around 2010
☐ …is a data storage solution designed for multi-datacenter deployments
☑ is a data collection pipeline with adapters for different data sources called '*Beats'
☑ …part of the ELK stack
☐ …a streaming system with exactly-once guarantees

7. Mark the correct claims about HDFS replication and distances:

☑ Distance 6 is replication to another datacenter
☐ Distance 2 is off-rack replication
☐ Distance 0 is replication in the same datacenter
☑ Distance 0 is replication on the same node
☑ Distance 2 is in-rack replication
☐ Distance 4 is replication to another datacenter

8. Mark the true claims about Apache Cassandra

☑ It is a key-value store.
☐ It is a real-time, stream mining solution.
☐ The term 'shard' is used in the context of Cassandra fragmentation.
☑ It uses a Gossip-based protocol for communication

9. The key phases of the MapReduce data analysis platform are:

☐ The 'shuffle' phase sorts, copies and merges the intermediate outputs of its preceding phase(s)
☐ The 'muffle' phase which select data according predefined probabilities
☐ The 'map' phase which operates on an input split of a dataset and never spills data to local disk
☑ The 'reduce' phase which operates on different keys
☑ The 'map' phase which operates on an input split of a dataset and might spill data to local disk

10. The MapReduce platform…

☑ …was open-sourced at one stage of its lifetime
☑ …was among the first to tackle distributed data analytics on a massive scale
☐ …was developed by LinkedIn
☑ …is considered outdated
☐ …was completely open-source from the start

11. Mark the solutions which do not natively support real-time stream mining:

☑ Apache Spark without its Streaming components
☐ Apache storm
☐ Apache spark streaming
☐ Apache Kafka

☑ cassandra

12. Storm is …

☑ .. a real time, distributed stream processing platform
☑ regarded as one of the first truly operational streaming system with non-batch analytics and strong correctness guarantees
□ a real time, distributed data storage platform
□ a real time, distributed visualization platform
□ a non-real-time, distributed stream processing platform

13. Mark the correct claims about Apache Kafka:

□ It is limited to processing-time windowing
☑ it is a data ingestion and processing framework
□ it is a good choice when considering long-term ata storage
☑ Provides strong consistency guarantees
□ Not open-source
☑ Originated from LinkedIn

14. Mark the correct claims about Spark:

□ Spark never commits the outputs of a processing stage to disk (neither SSD, nor regular hard drives)
□ Bosch is key contributor to Spark
☑ Spark owes its excellent performance to its restricted, distributed shared memory implementation
□ Google developed and used Spark until 2013-2014
☑ Spark was among the first distributed data analysis platforms to offer strong correctness guarantees
□ The term 'shard' is used in the context of Spark inter-control center synchronization

15. Flink is

☑ .. a unified stream and batch processing framework
□ .. a platform which receives inputs from 'sprouts' and outputs results into bolts
□ a platform in which transformations and actions are key concepts
□ … a streaming system in which actions trigger lazy evaluation
□ … a platform which supports distributed checkpoints

16. Candlestick charts are…

☑ often used in financial data analysis
□ useful in detecting positive (aka 'bullish') trends as a sequence of red candlesticks
□ useful in detecting negative (aka 'bullish') trends as a sequence of green candlesticks
☑ useful to represent multiple values for a period of time, e.g. 5-minute period
□ used since the ninth century BC

17. Mark the correct claims about choropleth graphs:

- They are usually tied to a geo map
- They rely on colors or shading

18. Mark the true claims about DataWrapper:

☐ Built by Google
☐ Allows use for license holders only
☑ It is primarily a web-based viz solution
☑ Requires minimum (visual) design skills
☑ used by a couple of large newspapers and journals
☐ Requires maximum coding skills
☐ Not used in Germany at all
☑ Requires minimum coding skill
☐ Used only in Germany

19. Mark the complex visualization types often used in general, open-source data analysis platforms:

☑ Dashboards
☑ Choropleth graphs
☐ Fuzz charts
☐ Yard-stick charts
☑ Histograms

20. A lineage graph is

☑ useful when there is a fault and parts of a distributed data analysis process have to be re-run
☑ a set of dependencies between intermediate results in cluster-computing platforms
☑ not used in the context of MapReduce
☑ relevant in the context of Spark's lazy evaluation
☑ highly relevant in the process of Spark transformations

21. Mark the correct claims about the Resilient Distributed Datasets (RDDs):

☐ RDDs contain only key-value pairs
☐ They are key components of Hadoop ecosystem
☑ They are lightweight fault-tolerant distributed memory implementation
☑ RDDs might contain different data types, not just key-value pairs
☐ Python objects are kept deserialized in RDDs
☐ Java objects are pickled in RDDs