Il testo riguarda la data science, la generazione di big data e le tecnologie associate. Riassumendo i punti chiave:

- La data science riguarda l'estrazione di informazioni e conoscenza dai dati utilizzando teorie e tecniche statistiche e informatiche.

- I data scientist sono esperti in queste aree con una forte comprensione del business e dei dati. Il loro obiettivo è trasformare le informazioni nascoste nei dati in vantaggi competitivi.

- La data science supera alcune limitazioni della business intelligence, come l'utilizzo di dati strutturati e non strutturati, processi di estrazione flessibili e l'uso di tecnologie Big Data e NoSQL.

- La data science è resa possibile dalla disponibilità di grandi dataset, nuove tecnologie per la manipolazione e gestione dei dati e tecniche di analisi avanzate.

- Le fonti di generazione di big data includono l'Internet of Things (IoT), i sistemi sanitari, le app, le assicurazioni e molte altre.

- Un data lake è un sistema o un repository per memorizzare i dati in diversi formati. Un data swamp è un data lake non gestito e non accessibile.

- Hadoop è un framework per il processing distribuito di grandi dataset.

- Esistono diverse architetture per il calcolo parallelo, come SMP, MPP, Grid Computing e HPC.

- La scalabilità è la capacità di un sistema di gestire una crescita della quantità di lavoro.

- Hadoop ha avuto un'evoluzione da Hadoop 1 a Hadoop 2, rendendolo più adatto per un'ampia gamma di scopi e supportando SQL e l'elaborazione online.

- Ci sono diverse distribuzioni di Hadoop, tra cui Cloudera, Hortonworks, MapR e IBM InfoSphere BigInsight.

- Hadoop può essere utilizzato in cloud per motivi di costo e completezza del marketplace.

- Hadoop semplifica la gestione e la programmazione di cluster distribuiti, ma non è efficiente quanto le architetture centralizzate o appliance. Tuttavia, può adattarsi a diversi scopi a prezzi contenuti.  
  
Il testo fornisce una comparazione tra Hadoop 1 e Hadoop 2, due versioni del framework di elaborazione distribuita Hadoop, insieme a una panoramica di alcune utility comuni utilizzate in Hadoop.

Hadoop 1 è stato rilasciato a novembre 2012 e supporta solo il modello di elaborazione MapReduce (MR). In questa versione, MapReduce si occupa sia delle elaborazioni che della gestione del cluster. La scalabilità è limitata a 4.000 nodi per cluster. Hadoop 1 utilizza il concetto di "slot", che può eseguire sia attività di map che di reduce. Un singolo Namenode gestisce l'intero namespace, ma questo può essere un punto di rottura (Single-Point-of-Failure, SPOF) e richiede un intervento manuale in caso di guasto. Le API MR sono compatibili in tutta la versione principale Hadoop1x, ma un guasto al Namenode influisce sull'intero stack. Inoltre, Hadoop 1 non supporta Microsoft Windows.

Hadoop 2, rilasciato quasi un anno dopo, consente di lavorare con MapReduce e altri modelli di computazione distribuita come Spark, Hama, Girap, MPI e HBase coprocessors. La gestione delle risorse del cluster viene effettuata tramite YARN (Yet Another Resource Negotiator), mentre le elaborazioni vengono eseguite utilizzando i diversi modelli di computazione. La scalabilità viene aumentata a 10.000 nodi per cluster. Hadoop 2 utilizza il concetto di "container" che può eseguire task generici. I server Namenode multipli gestiscono multipli namespace. Per superare il problema del SPOF, Hadoop 2 ha un Namenode in standby pronto a intervenire in caso di guasto e supporta il ripristino automatico. L'utilizzo di Hadoop 2 come piattaforma è esteso e può essere utilizzato per elaborazioni di eventi, streaming e operazioni in tempo reale. Hadoop 2 supporta anche Microsoft Windows.

Oltre a Hadoop, il testo menziona diverse utility comuni utilizzate in Hadoop:

- HDFS (Distributed File System): un file system distribuito che fornisce accesso ai dati con un elevato throughput.

- YARN (Yet Another Resource Negotiator): un framework per la pianificazione dei job e la gestione delle risorse del cluster.

- MapReduce: un sistema basato su YARN per l'elaborazione parallela di grandi insiemi di dati.

- Ambari: uno strumento web-based per fornire, gestire e monitorare i cluster Apache Hadoop, inclusi HDFS, MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig e Sqoop.

- ZooKeeper: un servizio di coordinazione ad alte prestazioni per applicazioni distribuite.

- Oozie: un scheduler di workflow per gestire i job di Apache Hadoop. I job di Oozie sono grafi aciclici direzionati (DAG) di azioni.

- Sqoop: uno strumento per il trasferimento massivo di dati tra Hadoop e datastore strutturati, come i database relazionali.

- Pig: una piattaforma per l'analisi di grandi insiemi di dati che utilizza un

linguaggio di alto livello chiamato Pig Latin.

- Hive: un meccanismo per strutturare i dati su HDFS e interrogarli utilizzando un linguaggio SQL-like chiamato HiveQL.

- HBase: un DBMS distribuito non relazionale che fornisce funzionalità simili a BigTable per Hadoop.

- Spark: un framework che migliora le prestazioni di elaborazione rispetto a MapReduce e supporta il machine learning e l'elaborazione in memoria.

- Storm: un sistema per la computazione distribuita in tempo reale.

- Giraph: un sistema grafico iterativo per l'elaborazione distribuita di grafi.

- HCatalog: un'astrazione relazionale dei dati su HDFS che semplifica l'accesso e la gestione dei dati.

- Phoenix: un layer di database relazionale basato su HBase che garantisce bassa latenza nelle query.