Chapter – 10
Overview of Future Skills and Artificial Intelligence

Part.1

Big Data Analytics- बिग डेटा शब्द का प्रयोग 1990 के दशक से किया जा रहा है, कुछ लोग इस शब्द को लोकप्रिय बनाने का श्रेय जॉन मैशी को देते हैं।
The term big data has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term.

डेटा एनालिटिक्स के जनक कौन हैं? (Who is the father of Big Data?)
John Tukey
In 1962, John Tukey described a field he called "data analysis", which resembles modern data science. 1962 में, जॉन टुके ने एक क्षेत्र का वर्णन किया जिसे उन्होंने "डेटा विश्लेषण" कहा, जो आधुनिक डेटा विज्ञान से मिलता जुलता है।

बिग डेटा एनालिटिक्स जानकारी को उजागर करने के लिए बड़े डेटा की जांच करने की अक्सर जटिल प्रक्रिया है - जैसे कि छिपे हुए पैटर्न, सहसंबंध, बाजार के रुझान और ग्राहक प्राथमिकताएं - जो संगठनों को सूचित व्यावसायिक निर्णय लेने में मदद कर सकती हैं।
Big data analytics is the often complex process of

examining big data to uncover information -- such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

व्यापक पैमाने पर, डेटा एनालिटिक्स प्रौद्योगिकियां और तकनीकें संगठनों को डेटा सेट का विश्लेषण करने और नई जानकारी इकट्ठा करने का एक तरीका प्रदान करती हैं। बिजनेस इंटेलिजेंस (बीआई) प्रश्न व्यवसाय संचालन और प्रदर्शन के बारे में बुनियादी सवालों के जवाब देते हैं।
On a broad scale, data analytics technologies and techniques give organizations a way to analyze data sets and gather new information. Business intelligence (BI) queries answer basic questions about business operations and performance.

बड़े डेटा विश्लेषण प्रक्रिया के चार चरण (Four Steps of Big Data Analysis Process)

डाटा विश्लेषण प्रक्रिया में सम्मिलित चरण इस प्रकार है-

1. डेटा पेशेवर विभिन्न स्रोतों से डेटा एकत्र करते हैं। अक्सर, यह अर्ध-संरचित और असंरचित डेटा का मिश्रण होता है। जबकि प्रत्येक संगठन अलग-अलग डेटा स्ट्रीम का उपयोग करेगा, कुछ सामान्य स्रोतों में शामिल हैं:

Data professionals collect data from a variety of different sources. Often, it is a mix of semistructured and unstructured data. While each organization will use different data streams, some common sources include:

- internet clickstream data;
- web server logs;
- cloud applications;
- mobile applications;
- social media content;
- text from customer emails and survey responses;
- mobile phone records; and
- machine data captured by sensors connected to the internet of things (IoT).

2. डेटा तैयार और संसाधित किया जाता है। डेटा को डेटा वेयरहाउस या डेटा लेक में एकत्र और संग्रहीत करने के बाद, डेटा पेशेवरों को विश्लेषणात्मक प्रश्नों के लिए डेटा को ठीक से व्यवस्थित, कॉन्फ़िगर और विभाजित करना होगा। संपूर्ण डेटा तैयारी और प्रसंस्करण विश्लेषणात्मक प्रश्नों से उच्च प्रदर्शन प्रदान करता है।
Data is prepared and processed. After data is collected and stored in a data warehouse or data lake, data professionals must organize, configure and partition the data properly for analytical queries. Thorough data preparation and processing makes for higher performance from analytical queries.

3. इसकी गुणवत्ता में सुधार के लिए डेटा को साफ़ किया जाता है। डेटा पेशेवर स्क्रिप्टिंग टूल या डेटा गुणवत्ता सॉफ़्टवेयर का उपयोग करके डेटा को साफ़ करते हैं। वे किसी भी त्रुटि या विसंगतियों, जैसे दोहराव या स्वरूपण गलतियों की तलाश करते हैं, और डेटा को व्यवस्थित और सुव्यवस्थित करते हैं।
Data is cleansed to improve its quality. Data professionals scrub the data using scripting tools or

data quality software. They look for any errors or inconsistencies, such as duplications or formatting mistakes, and organize and tidy up the data.

4. एकत्रित, संसाधित और साफ़ किए गए डेटा का एनालिटिक्स सॉफ़्टवेयर के साथ विश्लेषण किया जाता है। इसमें निम्न के लिए उपकरण शामिल हैं: The collected, processed and cleaned data is analyzed with analytics software. This includes tools for:

- ❖ डेटा माइनिंग (Data mining)- जो पैटर्न और रिश्तों की तलाश में डेटा सेट के माध्यम से छान-बीन करता है | which sifts through data sets in search of patterns and relationships
- ❖ पूर्वानुमानित विश्लेषण (Predictive analytics)- जो ग्राहक व्यवहार और अन्य भविष्य की कार्रवाइयों, परिदृश्यों और रुझानों का पूर्वानुमान लगाने के लिए मॉडल बनाता है | which builds models to forecast customer behavior and other future actions, scenarios and trends
- ❖ मशीन लर्निंग (Machine learning)- जो बड़े डेटा सेट का विश्लेषण करने के लिए विभिन्न एल्गोरिदम का उपयोग करता है | which taps various algorithms to analyze large data sets
- ❖ डीप लर्निंग (Deep Learning) - जो मशीन लर्निंग की एक अधिक उन्नत शाखा है | which is a more advanced offshoot of machine learning.
- ❖ टेक्स्ट माइनिंग और सांख्यिकीय विश्लेषण सॉफ्टवेयर (text mining and statistical analysis software)
- ❖ कृत्रिम बुद्धिमत्ता (एआई) (artificial intelligence (AI))

- ❖ मुख्यधारा बिजनेस इंटेलिजेंस सॉफ्टवेयर (mainstream business intelligence software)
- ❖ डेटा विज़ुअलाइज़ेशन उपकरण (data visualization tools)

प्रमुख बड़े डेटा विश्लेषण प्रौद्योगिकियाँ और उपकरण (Key big data analytics technologies and tools) - बड़े डेटा विश्लेषण प्रक्रियाओं का समर्थन करने के लिए कई अलग-अलग प्रकार के टूल और तकनीकों का उपयोग किया जाता है। बड़े डेटा विश्लेषण प्रक्रियाओं को सक्षम करने के लिए उपयोग की जाने वाली सामान्य तकनीकों और उपकरणों में शामिल हैं:
Many different types of tools and technologies are used to support big data analytics processes. Common technologies and tools used to enable big data analytics processes include:

- ❖ **Hadoop**-जो बड़े डेटा सेट को संग्रहीत और संसाधित करने के लिए एक खुला स्रोत ढांचा है। Hadoop बड़ी मात्रा में संरचित और असंरचित डेटा को संभाल सकता है।
  which is an open source framework for storing and processing big data sets. Hadoop can handle large amounts of structured and unstructured data.

  Note – सबसे बड़ा हंदूप क्लस्टर कौन सा है – फेसबुक

- ❖ **प्रिडिक्टिव एनालिटिक्स हार्डवेयर और सॉफ्टवेयर (Predictive analytics hardware and software)-** जो बड़ी मात्रा में जटिल डेटा को संसाधित करते हैं, और भविष्य की घटना के परिणामों के बारे में भविष्यवाणी करने के लिए मशीन लर्निंग और सांख्यिकीय एल्गोरिदम का उपयोग करते हैं। संगठन धोखाधड़ी का पता लगाने, विपणन, जोखिम

मूल्यांकन और संचालन के लिए पूर्वानुमानित विश्लेषण उपकरण का उपयोग करते हैं।
which process large amounts of complex data, and use machine learning and statistical algorithms to make predictions about future event outcomes. Organizations use predictive analytics tools for fraud detection, marketing, risk assessment and operations.

- **स्ट्रीम एनालिटिक्स टूल (Stream analytics tools)-** जिनका उपयोग बड़े डेटा को फ़िल्टर करने, एकत्र करने और विश्लेषण करने के लिए किया जाता है जिसे कई अलग-अलग प्रारूपों या प्लेटफार्मों में संग्रहीत किया जा सकता है।
which are used to filter, aggregate and analyze big data that may be stored in many different formats or platforms.
- **वितरित भंडारण डेटा (Distributed storage data)-** जिसे आम तौर पर एक गैर-संबंधपरक डेटाबेस पर दोहराया जाता है। यह स्वतंत्र नोड विफलताओं, खोए हुए या दूषित बड़े डेटा के विरुद्ध एक उपाय के रूप में, या कम-विलंबता पहुंच प्रदान करने के लिए हो सकता है।
which is replicated, generally on a non-relational database. This can be as a measure against independent node failures, lost or corrupted big data, or to provide low-latency access.
- NoSQL डेटाबेस (NoSQL databases)- जो गैर-संबंधपरक डेटा प्रबंधन प्रणालियाँ हैं जो वितरित डेटा के बड़े सेट के साथ काम करते समय उपयोगी होते हैं। उन्हें किसी निश्चित स्कीमा की आवश्यकता नहीं होती है, जो उन्हें कच्चे और असंरचित डेटा के लिए आदर्श बनाता है।

which are non-relational data management systems that are useful when working with large sets of distributed data. They do not require a fixed schema, which makes them ideal for raw and unstructured data.

- **एक डेटा वेयरहाउस (A data warehouse)-** जो एक रिपॉजिटरी है जो विभिन्न स्रोतों द्वारा एकत्र किए गए डेटा की बड़ी मात्रा को संग्रहीत करता है। डेटा वेयरहाउस आमतौर पर पूर्वनिर्धारित स्कीमा का उपयोग करके डेटा संग्रहीत करते हैं। which is a repository that stores large amounts of data collected by different sources. Data warehouses typically store data using predefined schemas.