# Tutorial Letter 102/3/2023

## Databases Systems IV

# IRM4723

## Year: 2023

## School of Computing

**IMPORTANT INFORMATION:**

This tutorial letter contains important information about your module.

# Content

## Special note on summative assessment

The format of the final (summative) assessment (examination) for 2023 is similar to that of 2022 for both the main and supplementary (Jan/Feb 2024) exams. This will be a multiple-choice questions (MCQs) only online non-venue-based assessment. As usual, the final assessment will be based on the ENTIRE syllabus and the study materials mentioned in tutorial letters 101 and 102 (Chapters 10 to 16 of the prescribed textbook and Tutorial letter 102). Do not study only the questions asked in the assignments as this will not be sufficient to pass the exam. Likewise, do not rely on past paper questions only.

This note is written in alignment with the statement below which has been approved by UNISA. I, therefore, request that you **do not request me to send you more information or scope with respect to the exam**. As mentioned in section 5.2 of Tutorial Letter 101, “The syllabus is covered by Chapters 9, 10, 11, 12, 13, 14, 15 and 16 of the prescribed textbook plus the content of Tutorial letter 102. “

**Demarcation or scoping of examinations and assessment**

“…academics were not to demarcate or scope specific work for examination purposes, but that examination questions should be based on the entire work covering the notional hours of the modules.”

# Data Management and Big Data

## 1. Introduction

In recent time, data have become a stream flowing into every area of the global economy. Many organizations produce volumes of transactional data that consist of trillions of bytes of information about their customers, suppliers and operations. A number of networked sensors that are embedded in devices such as smart phones, smart energy meters, automobiles and industrial machines create and communicate data in this century of the Internet of things.

Many organizations transact their businesses through interaction with individual thereby generating a tremendous amount of digital data. These data are created as a result of activities on social media sites, smartphones and others devices such as PCs and laptops. The growing volume of multimedia content is playing a major role in the exponential growth of big data.

Big data is referred to as the type of datasets which has capability beyond the ability of typical database software tools in capturing, storing, managing and analysing data. The term big data is not defined with regards to being larger than a certain number of terabytes because the sizes of datasets that qualifies as big data increases over time. Therefore the definition varies by sector based on the types of software tools that are available and sizes of the datasets that are commonly employed in a particular Industry. However, big data could be referred to as dataset that ranges from terabytes to multiple petabytes.

The recent growth of data usage is a global phenomenon. Many organizations regard this collection of information with a high level of suspicion with the notion that data flood is an intrusion of privacy. However, there has been strong evidence that big data is playing a significant role to both private commerce and the national economy. Big data enhances productivity and competitiveness thereby creating substantial economic surplus for consumers. For example, the US health sector employs big data effectively to drive efficiency and quality by reducing national health care expenditure by 8% (saving about $ 300 billion/per annum). In Europe, the government administration saves $ 149 billion in operational efficiency improvement by employing big data in their activities.

Big data has been creating value across several sectors of organizations and the global economy. It has created tremendous growth in innovation, productivity and new modes of competition. These trends have enabled consumers, organizations and economic sectors to exploit its potential. Although data have always been a key aspect of the impact of Information Communication Technology (ICT), the scale and scope of changes that are brought by big data have led to a visible acceleration in technology trends and convergence.

Organizations are using big data to create value while government are using big data to enhance efficiency and create the value for money on their offerings to the citizens at a time when there is limited resource allocation.

## 2. Definition and components of Big Data

Have a look at section 2.4 on page 39 of the text book. This section elaborated on the evolution of data management models from 1960s to present. It started from the first generation of file systems to the emerging models of NoSQL generation.

It has become an important need for an organization to derive usable business information (such as purchasing histories, behaviour patterns, customer preference, and browsing patterns) from masses of web data the organization have accumulated over the year. The loads of data can be a combination of structured and unstructured data from various sources.

Today’s organisation's databases are faced with the challenges of:

- Rapid data growth
- System performance and Scalability

So as an Information Technology (IT) Manager, you need to constantly balance and manage the rapidly growing data with shrinking budget. The need to manage rapid growth, performance, scalability and lower budget have prompted a phenomenon called **Big Data**.

**Big Data** is referred to as a complex and large set of structured and unstructured data that are incapable of being stored and processed by current technologies and infrastructures. It is a movement to find new and better ways to manage large amounts of web-generated data and derive business insights from it, while simultaneously providing high performance and scalability at a reasonable cost.

**The main components of Big Data are:**

1) **Volume:** This refers to quantity of data available from various sources
2) **Velocity:** Velocity is used in big data to depict the speed of data creation. The rate at which data changes.
3) **Variety:** Big Date can take different forms such as messages, updates, and social network images and it comes from different sources
4) **Veracity:** This relates to the trustworthiness of the large sets of data collected from various sources.

These four components are referred to as the four “V ‘ S” of Big Data.

The need of organizations with the challenge of Big Data cannot be resolved with the relational approach to databases.

Study page 51 section 2.4.6 for more reasons why relational approach to databases is not suitable for the needs of organisations with Big Data challenges.

Also, study section 2.4.7 (Page 52-54) “NoSQL Database” a term used for new generations of databases that addresses the specific challenges of Big Data era.

## 3. Challenges of Big Data

Organizations, national leaders and policy makers need to address some certain challenges so as to capture the full potential of big data. These challenges are:

- Shortage of the analytical and managerial individual that are experienced enough to make use of big data. For example, US have a shortage of 140,000 to 190,000 people with deep analytical skills and 1.5 million managers/analyst that could analyse big data and make suitable recommendations in their findings.
- Lack of adequate Infrastructure, incentives and competition that would support continued innovation.

Other Five managerial challenges companies need to address in order to reap the full benefits of big data or transition to big data include:

1. **Leadership**. Companies succeed in the big data era not simply because they have more or better data, but because they have leadership teams that set clear goals, define what success looks like, and ask the right questions. Big data’s power does not erase the need for vision or human insight. On the contrary, we still must have business leaders who can spot a great opportunity, understand how a market is developing, think creatively and propose truly novel offerings, articulate a compelling vision, persuade people to embrace it and work hard to realize it, and deal effectively with customers, employees, stockholders, and other stakeholders. The successful companies of the next decade will be the ones whose leaders can do all that while changing the way their organizations make many decisions.

2. **Talent management**. As data become cheaper, the complements to data become more valuable. Some of the most crucial of these are data scientists and other professionals skilled at working with large quantities of information. Statistics are important, but many of the key techniques for using big data are rarely taught in traditional statistics courses. Perhaps even more important are skills in cleaning and organizing large data sets; the new kinds of data rarely

come in structured formats. Visualization tools and techniques are also increasing in value. Along with the data scientists, a new generation of computer scientists are bringing about techniques for working with very large data sets.

3. **Technology.** The tools available to handle the volume, velocity, veracity, and variety of big data have improved greatly in recent years. In general, these technologies are not excessively expensive, and much of the software is open source. For example, Hadoop is one of the most commonly used frameworks. It combines commodity hardware with open-source software. It takes incoming streams of data and distributes them onto cheap disks; it also provides tools for analysing the data. However, these technologies do require a skill set that is new to most IT departments, which will need to work hard to integrate all the relevant internal and external sources of data. Although attention to technology isn't sufficient, it is always a necessary component of a big data strategy.

4. **Decision making**. An effective organization puts information and the relevant decision rights in the same location. In the big data era, information is created and transferred, and expertise is often not where it used to be. The clever leader will create an organization that is flexible enough to minimize the "not Getting Started You don't need to make enormous up-front investments in IT to use big data (unlike earlier generations of IT-enabled change). Here's one approach to building a capability from the ground up. (1) Pick a business unit to be the testing ground. It should have a qualified friendly leader backed up by a team of data scientists. (2) Challenge each key function to identify five business opportunities based on big data, each of which could be prototyped within five weeks by a team of no more than five people. (3) Implement a process for innovation that includes four steps: experimentation, measurement, sharing, and replication. (4) Keep in mind Joy's Law: "Most of the smartest people work for someone else." Open up some of your data sets and analytic challenges to interested parties across the internet and around the world. These were taken from Harvard Business Review October 2012. People who understand the problems need to be brought together with the right data, but also with the people who have problem-solving techniques that can effectively exploit them.

5. **Company culture.** The first question a data-driven organization asks itself is not "What do we think?" but "What do we know?" This requires a move away from acting solely on hunches and instinct. It also requires breaking a bad habit we've noticed in many organizations: pretending to be more data-driven than they actually are.

## 4. Advantages of Big Data

Big Data offer some of the following advantages:

- Creating transparency that making big data accessible to relevant stakeholders in a timely manner would create tremendous value. For example, accessibility to data across separate departments in organizations would reduce data redundancy and processing time. Data integration in engineering, R&D and manufacturing sectors would enable concurrency thereby reducing time to market entry and improve quality.
- Enabling experimentation to expose variability and improve performance when data are created and stored in digital form. Organizations collect accurate and detailed performance real-time data on products such as inventories, personnel sick days. The use of data helps to analyse variability in performance that occurs naturally or generated by experiments and to understand its root causes by enabling leaders to manage the performance to higher levels.
- Segmenting populations to customize actions by allowing organizations to create highly specific segmentations and tailor products and services to meet those needs. Big data techniques such as micro-segmentation of customers are used by marketing to target customers for promotions or advertisement.
- Replacing human decision making with automated algorithms by using sophisticated analytics that considerably improve decision making, minimize risks and unearth valuable insights that remain hidden. Organizations are making decisions by analysing datasets from customers, employees or sensors embedded in products.
- Innovating new business models, products and services through big data enables companies to create new products and services, enhance existing ones and invent new business models. Manufacturers are using data obtained from products to improve the development of next generation of products and create innovative after-sales service offerings.

## 5. Factors for Big Data Implementation

- Data policies: There should be a set of policies that are important in the organization when a large amount of data is digitised in organizational boundaries. These are privacy, security, intellectual property and liability. Privacy is very important particularly to customers whose personal data such as health and financial records but these data help to customer to determine the right medical treatment or find the appropriate financial product. There would be a trade-off between privacy and utility which individuals and societies would need to contend with. Data security is another major concern by protecting sensitive data to avoid data breaches that would expose consumer, corporate and national security information.
- Technology and techniques: Organizations would need to deploy new technologies and techniques. These technologies would depend on the data

maturity of the institution, legal systems and compatibility of data standard for better integration. The growth in technology innovation and techniques would assist organizations and individuals to integrate, analyse, visualize and consume the growing surge of big data.

- Organizational Challenge and talent: Lack of understanding of big data by leaders in organizations would lead to new entrants leveraging big data to compete against them. Many organizations do not have the resources to derive insights from big data thereby workflows and incentives are not structured so as to optimize the use of big data for better decisions and informed action.
- Access to data: Integration of data from multiple sources is important for organizational transformative opportunities therefor data access by a third party is often cumbersome and economic incentives are not aligned to encourage stakeholders to share data.

Organisations can improve their performance by exploiting enormous new flows of information. But for that to happen, they have to change their decision making cultures. Transition to using big data requires effective change management. **Change management** is one of the processes within service transition stage of ITIL framework. You will be introduced to change management in principal concepts of ITSM covered in IRM711.

## Meta-Data Management

### 6. Definition of Meta-Data Management

Metadata is referred to as data that used to describe other data. Metadata summarizes information about data by making finding and working with particular instances of data easier. For example, metadata files contain information about the author, date created, date modified and file size. Metadata are very useful for easy access through filtering to locate a specific document.

Metadata documents also contain files, images, videos and spreadsheets and web pages. Metadata can be created manually or through automated information processing. Manual processing tends to be more accurate by allowing the user to input information that are relevant or needed to help designate the file.

Automated metadata creation is elementary because it displays information such as file size, file extension, time of file creation and the author. The use of metadata on websites provides information of the page, keywords to the content, description of the page's content and the keywords linked to the content of the web page.

Metadata management is the process of ensuring data are associated with data assets to ensure that information can be integrated, accessed, shared, linked, analysed and maintained to best performance in an organization. It summarizes information about data so as to allow finding and working with instances of data easier. It provides the ability to filter through data for easier data asset location. Metadata occur when data is generated, acquired, added to, deleted and updated in the organization while metadata management is designed to ensure that metadata is added appropriately and that mechanisms are in place to optimize its effectiveness.

## 7. Advantages of metadata management

**Metadata management offers the following advantages to organisations:**

- Provide adequate maintenance of information across organization without depending on a particular employee's knowledge
- Provide efficient product and project delivery
- Provide less redundancy of effort and greater consistency across multiple instances of data.
- Provide appropriate re-use of data.

## Case study examples of how Managers uses big data

The PASSUR and Sears Holding examples illustrate the power of big data, which allows more accurate predictions, better decisions, and precise interventions, and can enable these things at seemingly limitless scale

### Improved Airline ETAs case study

Minutes matter in airports. So does accurate information about flight arrival times: If a plane lands before the ground staff is ready for it, the passengers and crew are effectively trapped, and if it shows up later than expected, the staff sits idle, driving up costs. So when a major U.S. airline learned from an internal study that about 10% of the flights into its major hub had at least a 10-minute gap between the estimated time of arrival and the actual arrival time—and 30% had a gap of at least five minutes—it decided to take action.

At the time, the airline was relying on the aviation industry's long-standing practice of using the ETAs provided by pilots. The pilots made these estimates during their final approach to the airport, when they had many other demands on their time and attention. In search of a better solution, the airline turned to PASSUR Aerospace, a provider of decision-support technologies for the aviation industry. In 2001 PASSUR began offering its own arrival estimates as a service called RightETA. It calculated these times by combining publicly available data about weather, flight schedules, and other factors with proprietary data the company itself collected, including feeds from a a network of passive radar stations it had installed near airports to gather data about every plane in the local sky.

PASSUR started with just a few of these installations, but by 2012 it had more than 155. Every 4.6 seconds it collects a wide range of information about every plane that it "sees." This yields a huge and constant flood of digital data. What's more, the company keeps all the data it has gathered over time, so it has an immense body of multidimensional information spanning more than a decade. This allows sophisticated analysis and pattern matching. RightETA essentially works by asking itself "What happened all the previous times a plane approached this airport under these conditions? When did it actually land?"

After switching to RightETA, the airline virtually eliminated gaps between estimated and actual arrival times. PASSUR believes that enabling an airline to know when its planes are going to land and plan accordingly is worth several million dollars a year at each airport. It's a simple formula: Using big data leads to better predictions, and better predictions yield better decisions.

**Sears Holding case study**

A couple of years ago, Sears Holdings came to the conclusion that it needed to generate greater value from the huge amounts of customer, product, and promotion data it collected from its Sears, Craftsman, and Lands' End brands. Obviously, it would be valuable to combine and make use of all these data to tailor promotions and other offerings to customers, and to personalize the offers to take advantage of local conditions. Valuable, but difficult: Sears required about eight weeks to generate personalized promotions, at which point many of them were no longer optimal for the company. It took so long mainly because the data required for these large-scale analyses were both voluminous and highly fragmented—housed in many databases and "data warehouses" maintained by the various brands.

In search of a faster, cheaper way to do its analytic work, Sears Holdings turned to the technologies and practices of big data. As one of first steps, it set up a Hadoop cluster. This is simply a group of inexpensive commodity servers whose activities are coordinated by an emerging software framework called Hadoop (named after a toy elephant in the household of Doug Cutting, one of its developers).

Sears started using the cluster to store incoming data from all its brands and to hold data from existing data warehouses. It then conducted analyses on the cluster directly, avoiding the time-consuming complexities of pulling data from various sources and combining them so that they can be analyzed. This change allowed the company to be much faster and more precise with its promotions. According to the company's CTO, Phil Shelley, the time needed to generate a comprehensive set of promotions dropped from eight weeks to one, and is still dropping. And these promotions are of higher quality, because they're more timely, more granular, and more personalized. Sears's Hadoop cluster stores and processes several petabytes of data at a fraction of the cost of a comparable standard data warehouse.

Shelley says he's surprised at how easy it has been to transition from old to new approaches to data management and high-performance analytics. Because skills and knowledge related to new data technologies were so rare in 2010, when Sears started the transition, it contracted some of the work to a company called Cloudera. But over time its old guard of IT and analytics professionals have become comfortable with the new tools and approaches.

The PASSUR and Sears Holding examples illustrate the power of big data, which allows more accurate predictions, better decisions, and precise interventions, and can enable these things at seemingly limitless scale. Big data has been used in supply chain management to understand why a carmaker's defect rates in their sudden increases, in customer service to continually scan and intervene in the health care practices of millions of people, in planning and forecasting to better anticipate online sales on the basis of a data set of product characteristics, and so on. Other industries and functions that use big data ranges, from finance to marketing to hotels and gaming, and from human resource management to machine repair.

## 8. Bibliography

- □ Lee, I., 2017. Big data: Dimensions, evolution, impacts, and challenges. Business Horizons, 60(3), pp.293-303.
- □ Manyika, J (2011). Big Data: The next frontier for innovation, competition, and productivity a growing torrent, McKinsey Global Institute.
- □ Michael C, Markus L, and Roger R (2010). The Internet of Things, McKinsey
  - o Global Institute.
- □ McAfee, A. and Brynjolfsso, E. (2012) Big Data: The Management Revolution.
  - o Harvard business review

This section is NOT examinable but as a person who has studied databases at a postgraduate level it would be good to have knowledge of some of the trends in databases/databases management/ databases management systems. See what follows.

**Part 1:**

For the following new trends in database management click the link given:

- Cloud-based DBMS
- Automation and DBMS
- Augmented DBMS
- Increased security
- In-memory databases
- Graph databases
- Open source DBMSs

- Databases-as-a-service

Link: https://tinyurl.com/5xbnb3a8

**Part 2:**

4 Top Trends in Database Management

https://www.datavail.com/blog/4-top-trends-db-management/