ICISSE 2023

Real-time big data analysis systems resulting from the Internet of Things (IoT)

Mohammed A. Makarem[1]

Muneef A. Razaz[1]

[1] *King Fahd university of petroleum and minerals(KFUPM)*

Outlines

ICISSE 2023

- **INTRODUCTION & Background**
  - Research Problem
  - Research Goals
- **Used Research Principles**
  - IoT and Smart Environment
  - Big Data
- **Frameworks for Big Data Processing: Spark and Hadoop**
  - Overview and components
  - Primary Distinctions Between Hadoop and Spark
- **The Literature review**
  - General Data
  - IoT Data
- **Methodology**
  - **Comparison Setup**
  - **Performance Testing**
- **Results**
- **Conclusion**

INTRODUCTION

ICISSE 2023

❑ **The Research Problem**:

- Increasing IoT Data Complexity: As the number of IoT devices grows, handling diverse and massive data in real-time becomes challenging.
- Importance of Real-time Analytics: Companies require efficient real-time analytics for timely decision-making.

❑ **Research Goals:**

- Objective: Evaluate data analysis systems for large-scale IoT data.
- Focus: Assess efficiency and suitability of Apache Spark and Apache Hadoop frameworks.

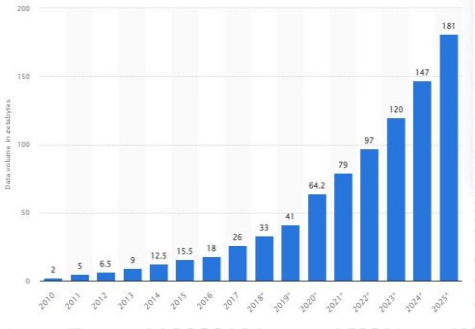

Figure 1. Data production from 2010 to 2025 in zettabytes [23].

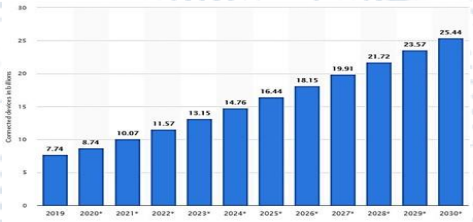

Fig. 2. . number of connected IoT devices from 2019 to 2030 [23].

Used Research Principles

- **IoT and Smart Environment**
  - ✓ **Definition:** IoT as an interconnected system providing unique identifiers for devices.
  - ✓ **Smart Environment:** Integration of devices for enhanced human comfort.
- **Big Data**
  - ✓ **Definition:** Large and complex datasets requiring advanced technologies.
  - ✓ **Importance:** Enhances decision-making processes across various fields.

Frameworks for Big Data Processing: Spark and Hadoop

ICISSE 2023

- **Hadoop**
  - ✓ **Overview:** Distributed processing using MapReduce algorithm.
  - ✓ **Components:** HDFS, MapReduce, YARN, and Hadoop Common library.
- **Spark**
  - ✓ **Overview:** High-performance data processing, supporting various workloads.
  - ✓ **Components:** Spark Core, Spark Streaming, Spark SQL, MLlib, and GraphX**.**

Primary Distinctions Between Hadoop and Spark

- **Performance Comparison:** Spark significantly faster, leveraging in-memory processing.
- **Cost and Resource Utilization:** Spark more cost-effective, utilizing less hardware.
- **Scalability:** Both frameworks scalable, but Spark excels in real-time processing.
- **Machine Learning:** Spark's in-memory calculations make it faster for ML algorithms.
- ***Hadoop vs. Spark Use Cases***
  - ✓ Hadoop Use Cases: Infrastructure setup on a budget, batch processing, historical data.
  - ✓ Spark Use Cases: Real-time streaming data analysis, fast results, machine learning.

Literature Review

ICISSE 2023

❑ **General Data**

✓ **Retail Sector:** Apache Spark used for Black Friday sales prediction.

✓ **Sentiment Analysis:** Hadoop employed for sentiment analysis on Twitter data.

❑ **IoT Data**

✓ **Smart Cities:** Framework leveraging MapReduce for IoT data processing.

✓ **Energy Management System (EMS):** IoT data used for optimizing energy consumption.

✓ **Smart Tourism:** TreSight system combining IoT and big data analytics in Trento, Italy.

✓ **Air Quality Prediction:** Spark and MLlib used for predicting air pollution from IoT sensors.

✓ **Healthcare Monitoring:** IoT and big data analytics for remote patient monitoring.

✓ **Industrial Sectors:** Hadoop and machine learning tools for fault prediction.

Methodology

ICISSE 2023

❑ Comparison Setup

- ✓ **Dataset:** Utilized IoT data from U.S. Environmental Protection Agency**.**
- ✓ **Pre-processing:** Used Pandas library for data cleaning and analysis.
- ✓ **Frameworks and Tools:** Spark 3.1.1 with Jupyter, Hadoop 3.1.1 with Windows 10 command prompt.
- ✓ **Performance Testing**

1. **Statistical Information:**
   - ✓ **Execution Time Comparison:** Spark outperformed Hadoop (67 seconds vs. 136.27 seconds).
2. **Machine Learning Operations:**
   - ✓ **Random Forest Classifier:** Spark performed significantly better (67 seconds vs. 240 seconds).
   - ✓ **Decision Tree Classifier**: Spark again outperformed Hadoop (58 seconds vs. 90 seconds).
3. **Data Flow:**
   - ✓ **Streaming Data Analysis:** Spark demonstrated real-time data processing capabilities.

Results

ICISSE 2023

❑ Execution Time

frameworks	Execution time/second
Hadoop (map-reduce)	136.27
Spark	67

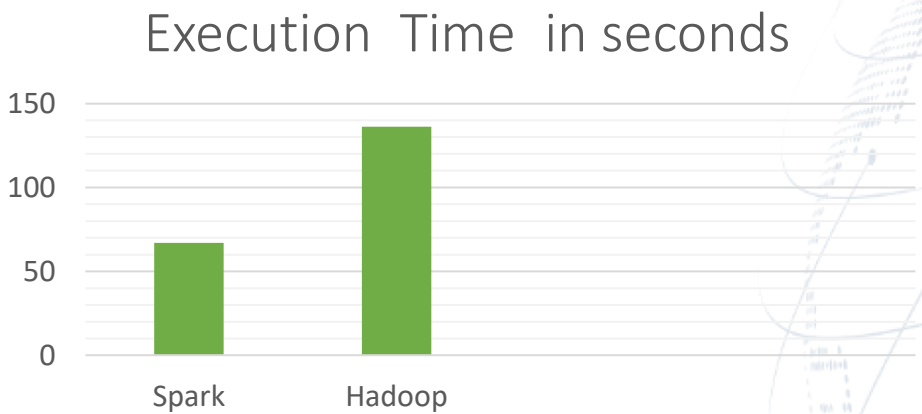

Figure 4.Time taken to perform the task.

Observations:

- ❖ Hadoop took twice the time of Spark for the same task.
- ❖ Hadoop: 2 minutes and a few seconds; Spark: 1 minute and a few seconds.

Results

ICISSE 2023

- **Second Experiment (Iterative Processes)**
  - **Random Forest Classifier Execution Time:**

Hadoop: 240 seconds

Spark: 67 seconds

  - **Decision Tree Classifier Execution Time:**

Hadoop: 90 seconds

Spark: 58 seconds

  - **Machine Learning Performance:**

Spark consistently outperformed Hadoop in both machine learning algorithms.

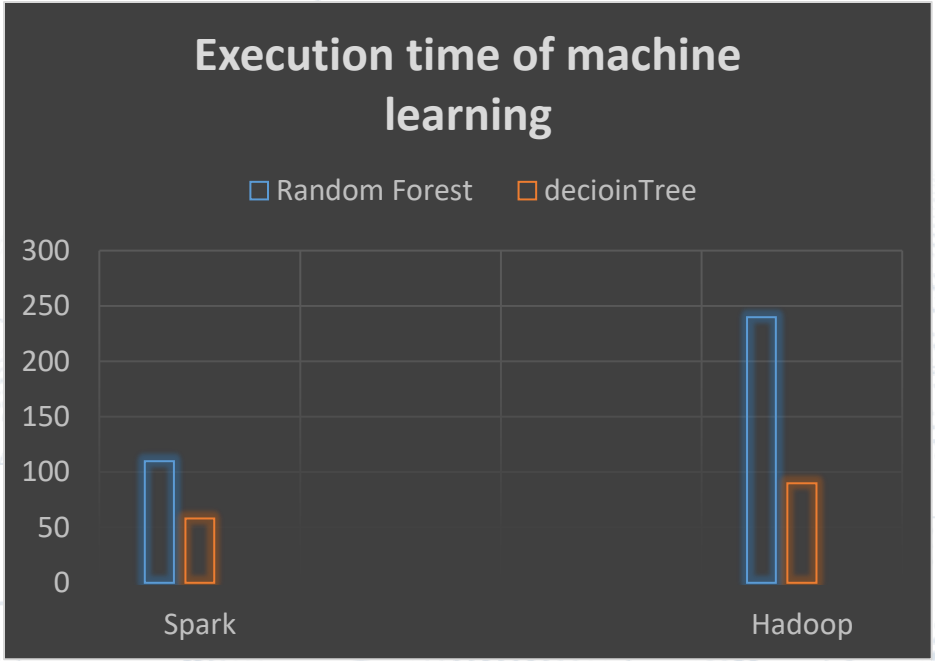

Figure 5. Execution time of machine learning methods.

Results

ICISSE 2023

❑ Third Experience (Data Flow)

- ***Real-time Streaming Data Processing:***

✓ Spark demonstrated speed and responsiveness.

✓ Live Chart showcased Spark's ability to handle streaming data.

✓ Spark's ability to operate solely in memory without resorting to hard disk usage contributed to its speed advantage.

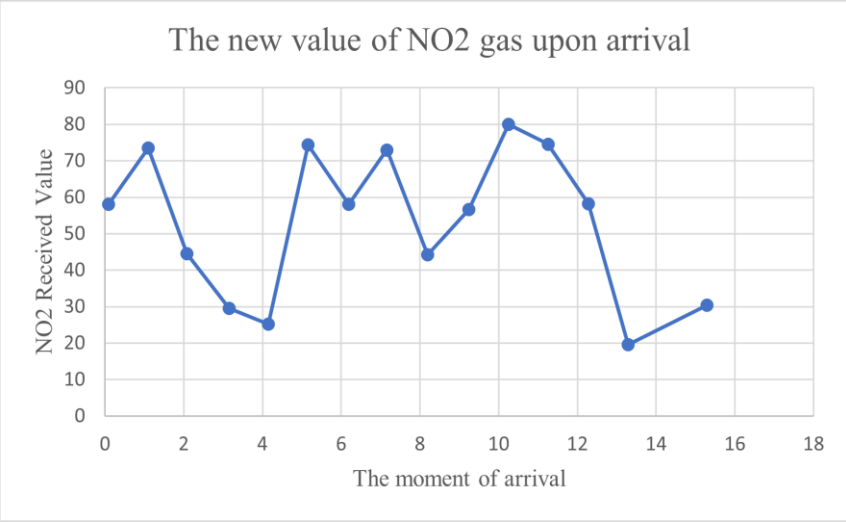

Figure 6.The flowing NO2 gas values.

Conclusion

ICISSE 2023

- ***Framework Suitability for IoT Data:***
  - ***Spark's Advantages:***
    - *Faster execution time.*
    - *Efficient resource utilization.*
    - *Superior performance in iterative processes.*
    - *Real-time data processing capabilities.*
    - *Cost-effective.*
- ***Implications for IoT Data Analysis:***
  - ***Significance:*** *As IoT data complexity grows, appropriate big data frameworks become crucial.*
  - ***Recommendation:*** *Spark identified as the more appropriate framework for large-scale streaming IoT data analysis.*
- ***Future Implications:***
- ***Continued Growth of IoT Data:*** *As IoT data volume and complexity increase, the choice of an efficient framework becomes paramount.*
- ***Research and Development:*** *Ongoing studies should explore advancements in big data frameworks to address evolving IoT data needs.*

Recommendations & Future work

ICISSE 2023

- **Adoption of Apache Spark:**
  - ***For Real-time Analytics:*** *Spark's in-memory processing and streaming capabilities make it ideal.*
  - ***Machine Learning Applications:*** *Superior performance in machine learning tasks.*

- ***Consideration of Specific Use Cases:***
  - ***Hadoop:*** *Batch processing, historical and archival data analysis.*
  - ***Spark:*** *Real-time streaming data analysis, machine learning applications..*
- **Future Work**
  - *Enhancements and Innovations:*
  - *Frameworks Evolution: Continuous improvement and innovation in both Hadoop and Spark.*
  - *Integration of Advanced Technologies: Exploration of emerging technologies for enhanced IoT data processing.*

Thank you for listening

Any Questions?

References

[37] S. A. Shah, D. Z. Seker, S. Hameed, and D. Draheim, “The rising role of big data analytics and IoT in disaster management: Recent advances, taxonomy and prospects,” *IEEE Access*, vol. 7, pp. 54595–54614, 2019, doi: 10.1109/ACCESS.2019.2913340.

[38] A. Aryal, Y. Liao, P. Nattuthurai, and B. Li, “The emerging big data analytics and IoT in supply chain management: a systematic review,” *Supply Chain Manag.*, vol. 25, no. 2, pp. 141–156, 2020, doi: 10.1108/SCM-03-2018-0149.

[39] M. Ge, H. Bangui, and B. Buhnova, “Big Data for Internet of Things: A Survey,” *Futur. Gener. Comput. Syst.*, vol. 87, pp. 601–614, 2018, doi: 10.1016/j.future.2018.04.053.