By: **Nadeen Ali** - 22110361

**Big Data Analytics & Visualization for Health Insights: Consulting Labs**

Course: Big Data Analytics & Visualization – Spring 2024/2025

Course instructor: Wa`ed Alsawareah

Table of Contents

[**1 Big Data Concept & DDDM in Healthcare** 3](#_Toc201182551)

[**1.1** **Advantages & Challenges of Big Healthcare Datasets** 3](#_Toc201182552)

[Advantages 3](#_Toc201182553)

[Challenges 3](#_Toc201182554)

[**1.2** **Predicting the Impact of Data Use on Consulting Labs** 4](#_Toc201182555)

[**2 Statistical & Graphical Techniques** 4](#_Toc201182556)

[**2.1 Merging, Cleaning, & Preparing Data** 4](#_Toc201182557)

[**2.2 Statistical Models & Graphical Techniques Used** 4](#_Toc201182558)

[Scenario 1 – Top 10 Requested Tests 4](#_Toc201182559)

[Scenario 2 – Monthly & Seasonal Trends of Requested Tests 6](#_Toc201182560)

[**3 MapReduce with Hadoop** 9](#_Toc201182561)

[**3.1 Word Count for Requested Test Names** 9](#_Toc201182562)

[**3.2 Data Preparation Methods & Tools Assessment** 14](#_Toc201182563)

[**4 Professional Challenges for Data Specialists in Healthcare** 15](#_Toc201182564)

[**4.1 Roles & Responsibilities of Data Specialists** 15](#_Toc201182565)

[**4.2 Data Compliance Strategies** 15](#_Toc201182566)

[**4.3 Real World Impact** 15](#_Toc201182567)

[**References** 16](#_Toc201182568)

# **1 Big Data Concept & DDDM in Healthcare**

## **Advantages & Challenges of Big Healthcare Datasets**

### Advantages

Because healthcare is a *very* sensitive field, dealing with real people (patients) and their problems (illnesses), it is crucial to ensure that a large amount of data is collected to gain reliable, accurate, and rich insights and information. This ensures that the data is diverse, covering a wide range of patients with different backgrounds, history, diseases, test results, etc., reducing potential bias in the data guarantee that analytics appropriately reflect the complexity of the population. This variation enables more accurate diagnosis and more customized medical care using predictive analytics.

In addition to improving patient outcomes, having access to large and high quality medical data helps healthcare and clinical companies implement data-driven decision making (DDDM). This helps them identify trends and patterns in the data such as the most requested tests, seasonal increases in illness, or test result progression. These insights prove to be extremely useful, making it easier to make decision that are well informed and backed up by *accurate* and *reliable* data, aiding in staff shifts and control, inventory management, resource allocation, etc. So, utilizing big data enhances the precision of decision making, allowing healthcare companies to promote clinical excellence.

### Challenges

But, of course, everything comes with a cost or a risk, and working with large healthcare datasets is *no* exception. The following include some potential risks healthcare companies must attend to:

1. Protecting data security and patient privacy *especially* by following strict regulations such as the GDPR and HIPPA.
2. Ensuring clean, healthy, and accurate data as integrating it from *many* different sources (patient history, doctor notes, lab results, etc.) can lead to missing data, data inconsistency, incorrect format, etc.
3. Managing technical costs as storing *massive* amounts of data such as healthcare is in need of incredibly expensive and scalable systems to maintain data storage.
4. Preserving quality of data over time due to the continuous data streaming which can easily result in mistakes, thus ensuring all data is in the same format and up to date with no duplicates, inconsistencies, or missing information.

## **Predicting the Impact of Data Use on Consulting Labs**

This usage of advanced data analytics will *greatly* benefit Consulting Labs, enhancing their services and customer experience.

1. **Patients:** Enables more accurate diagnoses, allowing quicker more precise and personalized treatment plans that will enhance patient health. Advanced analytics also allows monitoring a patients’ health over data (from visits, results, etc.) allowing early issue identification and customized care. Patients will also benefit from shorter wait times and better care.
2. **Management:** Management will gain a better understanding of operational trends, resource utilization, and test demand trends, allowing them to make data driven decisions regarding healthcare services to better suit the clinics’ current needs whether it be from the staff or patients. Some examples may include being better prepared for peak times, buying more of a certain medication or product, etc.
3. **Regulators:** Because Consulting Labs will be using advanced data analytics, their data will be cleaned, up to date, and well documented, thus making it easier for regulators to monitor their work and regulation.

# **2 Statistical & Graphical Techniques**

## **2.1 Merging, Cleaning, & Preparing Data**

The data was merged, cleaned, and prepared using PySpark (Apache Spark’s Python API).

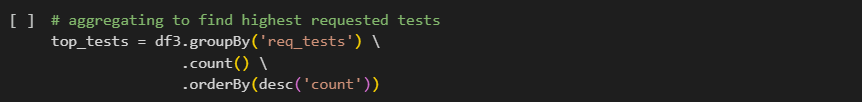
## **2.2 Statistical Models & Graphical Techniques Used**

After the data was fully preprocessed, different statistical and graphical techniques were implemented to further examine the data and uncover patterns.

### Scenario 1 – Top 10 Requested Tests

To find the top 10 most requested tests, here are the following steps:

1. **Groupby:** The data was sorted by grouped by column ‘req\_tests’ (requested tests) to count how many times each test was requested, ordered by the count.



1. **Print Result:** The 10 results were then printed

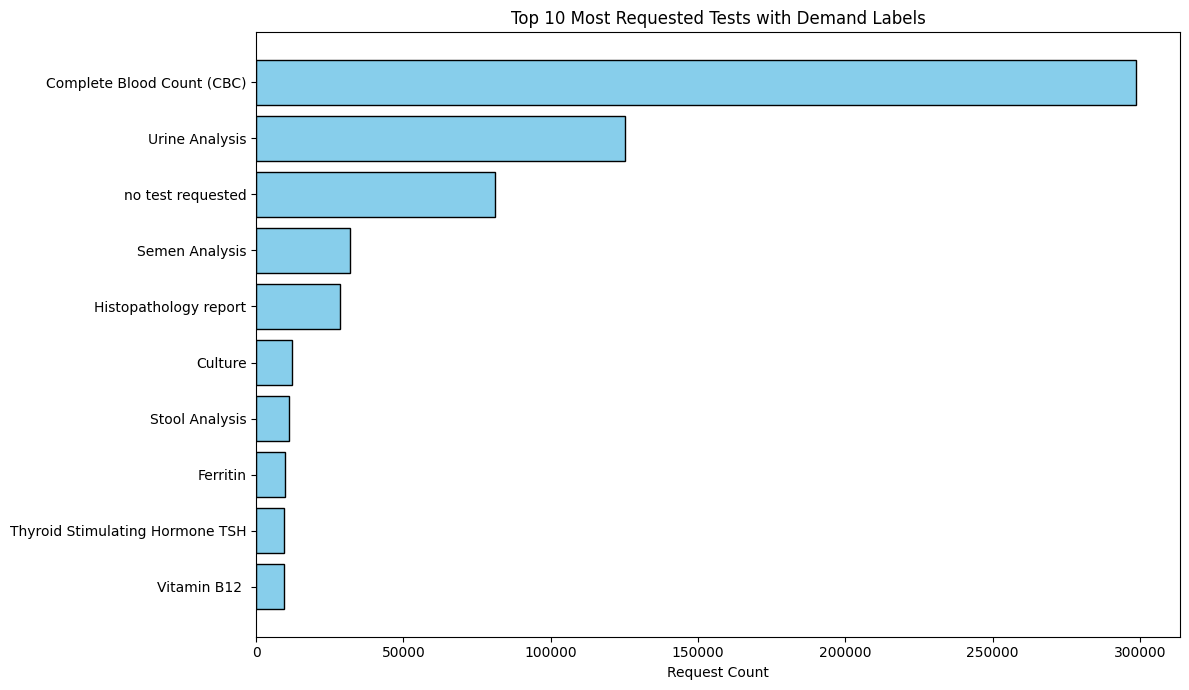
A screenshot of a computer

AI-generated content may be incorrect.

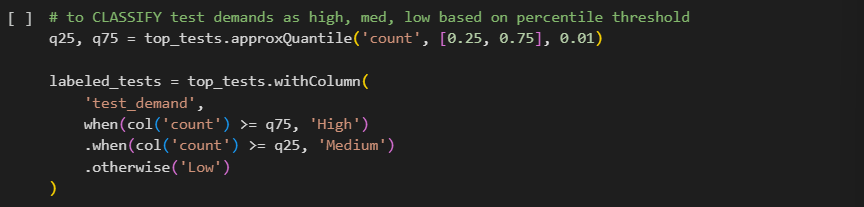
1. **Prepare to Plot:** The 10 requested tests were then saved in Pandas for easy plot access



1. **Bar Graph:** Using a *bar* graph, I visualized the top 10 most requested tests. We can visually see that the Completed Blood Count (CBC) test was by *far* the most requested followed by Urine Analysis. The 3rd largest is no test requested, showing that many patients who visit Consulting Labs do not get any tests requested to take.



1. **Demand Classification:** Using the top tests count, I calculated a threshold based on the 25th and 75th percentile to classify requested tests as High (greater than or equal to the 75th percentile), Medium (between 25th & 75th percentile), *and* Low (less than 25th percentile).



1. **Adding Feature:** This classification label was added to the original data

A computer screen shot of a black screen

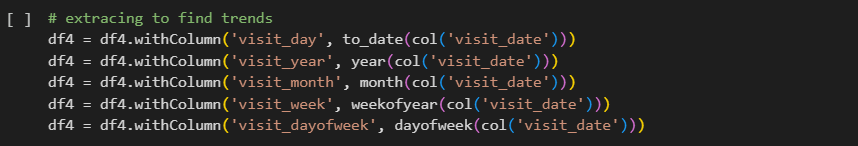
AI-generated content may be incorrect.

**Interpretation and Results:** Consulting labs can quickly determine which tests are most crucial by grouping them according to demand (high, medium, and low). The lab may prioritize resources, keep the required supplies on hand, and guarantee that staff members are properly trained for the most important tests by concentrating on the most requested tests. Better patient care and greater efficiency result from this.

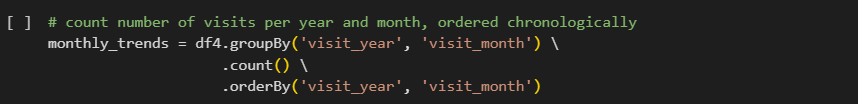
### Scenario 2 – Monthly & Seasonal Trends of Requested Tests

To analyze how test requests changed over time, these steps were followed:

1. **Feature Engineering:** Extracting visit day, month, year, week, *and* day of week from the ‘visit\_date’ column to track trends at different times.



1. **Monthly Aggregation:** The number of request tests was counted per month, then sorted chronologically to show how demand changed in order from beginning to the end.

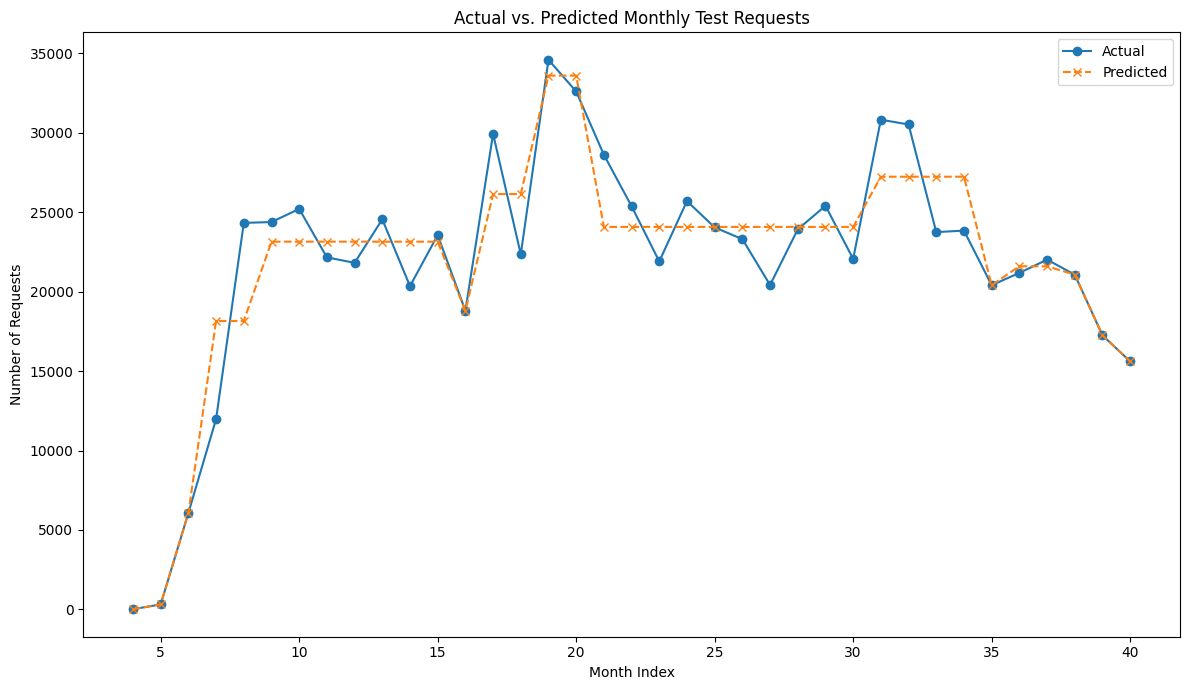


1. **Line Chart:** To visualize the changes in the trend over the months and years, a line graph was plotted to help highlight the trend overtime. As we can see, there is *usually* a peak in summer & fall months, and a drop during the spring.

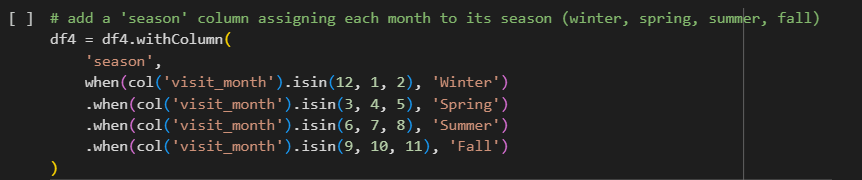
A graph with blue lines and numbers

AI-generated content may be incorrect.

1. **Predictive Modelling (Regression):** Using the Decision Tree Regressor, the count of requested tests per month were predicted and then plotted against the *actual* values to assess the model’s accuracy visually. As we can see in the figure below, the model follows the pattern quite well, capturing the overall peaks and drops of requested tests over time.



1. **Seasonal Analysis:** A new feature called ‘season’ was added to the data, classifying the requested test and its data requested based on its season.



1. **Bar Graph:** Lastly, the count of requested tests *per season* were visualized in a bar graph to make it easier to detect the busiest seasons. As expected from the line charts earlier, summer and fall see the *most* requested tests, with winter coming close afterwards and spring dropping decently.

A graph of blue rectangular shapes

AI-generated content may be incorrect.

**Interpretation and Results:** The demand for tests rises in the summer and fall and falls in the spring, so by determining the most popular times for requests, Consulting Lab can prepare ahead of time and guarantee they have enough staff and supplies during busy times. This makes the lab run more smoothly and guarantees timely and reliable patient care. Also, the lab can anticipate these trends and prepare for busy months by using predictive analysis.

# **3 MapReduce with Hadoop**

## **3.1 Word Count for Requested Test Names**

To implement MapReduce on my data as a word count task per requested tests, O extracted the ‘req\_tests’ column as a text file and ran a MapReduce word count job on it. The mapper and reducer codes are submitted as well as the text file, output, and steps to reach the output. The following are screenshots of the command prompt script to reach the output:

1. **Copying Files:** Needed files (clinical\_mapper.py, clinical\_reducer.py, *and* req\_tests.txt) are successfully copied to the Hadoop Docker container to execute.

A screenshot of a computer program

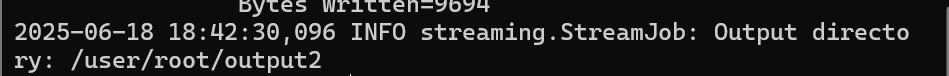
AI-generated content may be incorrect.

1. **MapReduce:** Executing the MapReduce word count task.

A screen shot of a computer program

AI-generated content may be incorrect.

1. **MapReduce Success:** Output direction is created at /user/root/output2



1. **Reading Output:** Displaying the final MapReduce word count output for requested tests

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

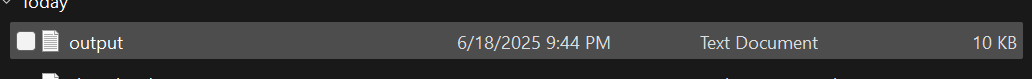
1. **Saving Output:** Saving MapReduce outputs to txt file and copying into local folder

A black screen with white text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.



## **3.2 Data Preparation Methods & Tools Assessment**

To prepare the data, I used **PySpark** to merge all 11 excel sheets as well as fully cleaning it, extracting features from it (feature engineering), and applying advanced aggregations methods to plot important graphs for visual aid. The data is *huge*, starting with ~300k rows and ending up with ~800k rows (due to exploding multi-line results such as CBC into multiple rows), thus the usage of PySpark is key as it enables fast processing of *many* rows, gives access to advanced aggregation methods, *and* ease of new feature creation to create a rich data for meaningful insights and plots.

As mentioned, I created plots to aid with visuals using **matplotlib**, making it easy to plot bar graphs and line charts for easy comparisons to highlight important patterns and trends in the data that otherwise could be missed if not visualized. These plots *also* helped non-technical peoples such as management to quickly understand important insights needed to better understand what Consulting Labs needs.

Lastly, the usage of **Hadoop MapReduce** allows for scalable manipulation and summarization of large data, making it easy and fast to extract key information such as my task of requested test counts.

So, all in all, Consulting Labs is able to make better business decisions due to the insights that are produced such as anticipating peak times whether monthly or seasonally to prevent being understaffed or underequipped with needed supplies and treatments for example and determining which tests require more staffing to monitor. This leads to better resource allocation, time management, and thus better patient care and improved regulatory compliance, easing Consulting Labs' operations short and long term.

# **4 Professional Challenges for Data Specialists in Healthcare**

## **4.1 Roles & Responsibilities of Data Specialists**

To collect, clean, and handle vast amounts of sensitive patient data, the healthcare sector needs data specialists. Data specialists in the healthcare sector also ensure data quality, integrating data from many sources, and creating reports and decision making visualizations. While conducting their roles and duties, they *must* maintain compliance with industry regulations, keeping patient data safe, secure, and private and ensuring compliance to regulations [1][2].

## **4.2 Data Compliance Strategies**

To ensure data compliance, data specialists use strategies including audits, data encryption, and strict access controls to ensure their work is constantly in compliance with regulations. Some of these regulations include HIPA or GDPR which aim to keep patient data private and protected. Another strategy include training teams to help keep them aware of *any* security risks and ensure they are aware of the needed strategies to implement to comply with regulations. However, complying with regulations and data privacy comes with challenges such as:

1. **Ethical**

* Receiving *clear* consent from patients to collect, store, and store their data
* Providing transparency and documentation on how the data is collected, handled, *and* used.
* Avoiding bias and discrimination in data analysis to ensure fairness

1. **Legal**

* Complying with industry regulations (such as HIPPA and GDPR as mentioned earlier)
* Adhering to laws such as anti-discrimination laws
* Protecting intellectual property

1. **Security**

* Guarding data against breaches, unauthorized access, attacks, etc.
* Using encryption, access controls, and other strong security measures to ensure sensitive data is kept safe.

## **4.3 Real World Impact**

Healthcare companies are heavily impacted by the issues and challenges that data professionals encounter. Any data security or privacy violation can easily result losing the trust of patients and risking legal repercussions. Thus, decisions based on inaccurate or biased information can have a negative influence on Consulting Labs. Thus, data specialists in Consulting Labs *must* ensure the data they are working with is not biased, represents patients as is, is formatted correctly with no missing or incorrect information, and decisions are made from *reliable and accurate* predictive analysis methods. This increases the organization's overall efficiency, enhances service quality, and fosters trust with patients and regulators.

# **References**

1. Subrahmanya, S.V.G., Shetty, D.K., Patil, V., Hameed, B.M.Z., Paul, R., Smriti, K., Naik, N. and Somani, B.K. (2021). The role of data science in healthcare advancements: applications, benefits, and future prospects. Irish Journal of Medical Science, [online] 191(4). doi:https://doi.org/10.1007/s11845-021-02730-z.
2. California Society for Respiratory Care. (2021). What does a Clinical Data Specialist do? Career Overview, Roles, Jobs | CSRC. [online] Available at: <https://jobs.csrc.org/career/clinical-data-specialist>.
3. CDC (2024). Health Insurance Portability and Accountability Act of 1996 (HIPAA). [online] Public Health Law. Available at: <https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html>.
4. European Union (2018). General Data Protection Regulation (GDPR). [online] General Data Protection Regulation (GDPR). Available at: <https://gdpr-info.eu/>.