# MINI PROJECT
# BIG DATA SYSTEM DESIGN

INSTRUCTION:

- Form a group of 7-9 people and register for the topic assignment at
- Design big data system on specified topic
- Make a video presentation 4-7 minutes long that includes:
- Problem statement
    - System design diagram (Data flow, components/software used, etc.)
    - Design Justification (Why do you design the system this way?)
    - Benefits of this design choice
    - Other suggestion
- Deadline on LEB2

**MEMBER:**


| | | |
|---|---|---|
| NAME : KHEMMANAT | BOONYACHEVITANON | ID : 64070503405 |
| NAME : JUMPONPATHA | CHAIMONGKONROJNA | ID : 64070503408 |
| NAME: THAMMARAT | DUJIMAYOON | ID: 64070503425 |
| NAME : PHANASORN | SRISAYAM | ID : 64070503436 |
| NAME : PHURICHAYA | JINTHANAWONG | ID : 64070503445 |
| NAME : WARUT | WANNASERT | ID : 64070503448 |
| NAME : SUEKSIT | VACHIRAKUMTHORN | ID : 64070503450 |
| NAME : INTOUCH | YUSO | ID : 64070503460 |
| NAME : SAWATSAKRON | MAHARUANKWAN | ID : 64070503475 |

## TOPIC 3: KEEPING A PROMISES

Stark Industry is planning to distribute a small model of Arc Reactor for household use. Chairperson promises to the users with 24/7 availability. So, they have to monitor for the remaining lifetime of the reactor and send a new one for replacement before failure occurs. In the initial phase, 1000 units of the reactor have been tested and the results are stored on the main database of the company. In the upcoming public release of the reactor, the marketing team expects over 1000000 units will be sold. You, as the maintenance team, decide to use survival analysis to predict the remaining lifetime of the reactor. They have to do this daily to ensure minimum failure. Arc reactor comes with the feature to send their current condition to the headquarter. So, they will have to design a new system to aggregate the data and predict the remaining lifetime of each unit. In summary, you will have to design a maintenance system that:

- Capable of aggregating data from over 1000000 units.
- Predicts remaining lifetime of each unit using Survival Analysis. Predicts daily.
- Update the survival analysis parameters from collected data every week.
- Generate the report of the maintenance process to the chairperson every week.

# Data-Flow Diagram

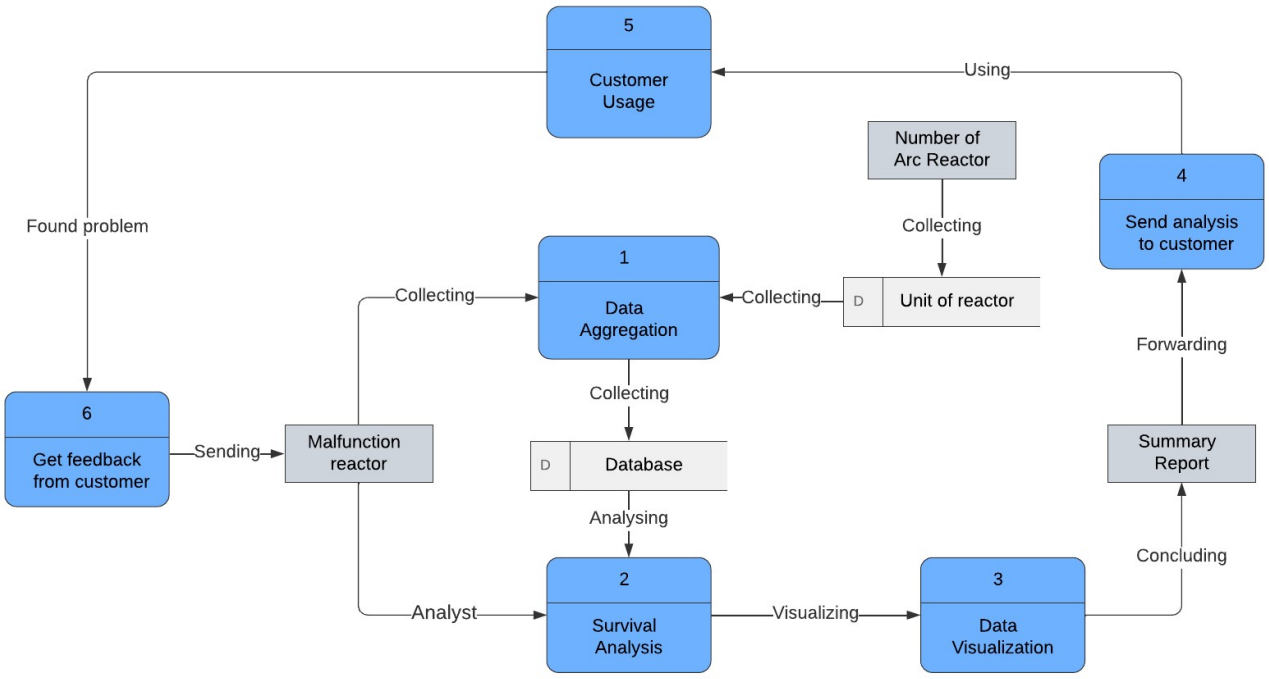


**Figure 1.** The data-flow diagram of Stark Industry.

# Data flow diagram explanation

1. **Data Collecting Module**
- Utilizes IoT devices within the reactors for data transmission.
- Data includes current condition, usage patterns, and any failure incidents.

2. **Data Aggregation Module**
- Aggregates and preprocesses the collected data from all units.
- Utilizes cloud-based storage (AWS S3, Azure Blob Storage) for scalable storage.

3. **Survival Analysis Module**
- Utilizes survival analysis algorithms (Kaplan-Meier, Cox Proportional Hazards) implemented in Python/R
- Incorporates collected data to update and refine survival analysis parameters.
- Runs daily predictions for each unit.

4. **Parameter Update Scheduler**
- Updates survival analysis parameters weekly based on aggregated data.
- Utilizes machine learning models (Bayesian methods) to update parameters.

5. **Data Visualization Module**
- Generates weekly maintenance reports using visualization tools (e.g., Tableau, Power BI) for clear and concise reporting.
- Includes information on predicted remaining lifetime for each unit, any units approaching failure, and recommendations for replacement.
- Sends reports to the Chairperson.

6. **Backup and Recovery**
- Implements backup mechanisms to prevent data loss.
- Defines a recovery plan in case of system failures.

# Design Justification

**IoT Devices for Data Collection:** These devices are embedded within each Arc Reactor and are connected through the Internet of Things (IoT). They serve the purpose of continuously collecting data in real-time. By doing so, they enable the system to monitor the reactors accurately and promptly. This real-time data transmission is crucial for keeping track of the performance, health, and operational status of each unit.

**Cloud-Based Storage:** All the data collected from these IoT devices is stored in cloud storage. This choice is made because cloud storage offers scalability and efficiency in handling a massive amount of data. With potentially millions of units generating data, using cloud storage allows for seamless management, accessibility, and organization of this data. It also facilitates easy retrieval and analysis, regardless of the volume.

**Survival Analysis Engine:** Survival analysis is a statistical method used to estimate the remaining lifespan or time-to-failure of units or systems. It's being employed to predict the remaining lifetimes of Arc Reactors. The data used for this analysis includes information from operational units. By implementing this analysis using programming languages like Python or R, a module is created to report daily on the expected lifetimes. This aids in proactive maintenance planning, allowing for intervention before potential failures.

**Parameter Update Module:** The predictive model used for monitoring the reactors requires regular updates to remain accurate and adaptable to changing conditions. Updating parameters weekly ensures that the model remains functional. This proactive approach helps maintain the accuracy of predictions and ensures the system can adapt to variations or changes in the reactors' operations.

**Reporting System:** These reports contain valuable insights derived from the collected data. They showcase maintenance trends, patterns, and any anomalies detected within the reactor units. By providing this information regularly, it promotes transparency and informed decision-making regarding strategic maintenance planning or other operational adjustments.

## Benefits of Design Choices

**Real-Time Monitoring:** IoT devices enable constant data transmission, facilitating real-time monitoring of unit conditions.

**Scalability:** Cloud-based storage ensures the system can handle data from a vast number of units without compromising performance.

**Predictive Maintenance:** Survival analysis predicts remaining lifetimes, enabling proactive replacement before failure occurs, reducing downtime and potential hazards.

**Adaptability:** Weekly parameter updates ensure the predictive model remains accurate, even with changing operational conditions

# Main Substance

## Survival Analysis:

Survival analysis is a set of statistical approaches used to determine the time it takes for an event of interest to occur. We use survival analysis to study the time until some event of interest occurs. Time is usually measured in years, months, weeks, days, and other time measuring units. The event of interest could be anything of interest. It could be an actual death, a birth, a retirement, along with others.

Survival analysis, also known as time-to-event analysis, is a branch of statistics that studies the amount of time it takes before a particular event of interest occurs.

- Survival analysis is a branch of statistics that studies how long it takes for certain instances to occur.
- It was initially developed in biomedical sciences to understand the onset of certain diseases but is now used in engineering, insurance, and other disciplines.
- Analysts at life insurance companies use survival analysis to estimate the likelihood of death at different ages, with health factors taken into account.
- This information is used to estimate the probability of a policyholder outliving their policy, which, in turn, influences insurance premiums.

### Advantages and Disadvantages of Survival Analysis

There are other more common statistical methods that may shed some light on how long it could take something to happen. For example, regression analysis, which is commonly used to determine how specific factors such as the price of a commodity or interest rates influence the price movement of an asset, might help predict survival times and is a straightforward calculation.

The problem is that linear regression often makes use of both positive and negative numbers, whereas survival analysis deals with time, which is strictly positive. More importantly, linear regression is not able to account for censoring, meaning survival data that is not complete for various reasons. This is especially true of right-censoring, or the subject that has not yet experienced the expected event during the studied time period.

The main benefit of survival analysis is that it can better tackle the issue of censoring as its main variable, other than time, addresses whether the expected event happened or not. For this reason, it is perhaps the technique

best-suited to answering time-to-event questions in multiple industries and disciplines.

Survival analysis is used in a variety of field such as:

- Cancer studies for patients survival time analyses.
- Sociology for “event-history analysis.”
- In Engineering for “failure-time analysis.”
- Time until product failure.
- Time until a warranty claim.
- Time until a process reaches a critical level.
- Time from initial sales contact to a sale.
- Time from employee hire to either termination or quit.
- Time from a salesperson to their first sale.

**Survival time and type of events in cancer studies**

Survival Time is usually referred to as an amount of time until when a subject is alive or actively participates in a survey.

There are three main types of events in survival analysis:

1. Relapse: Relapse is defined as a deterioration in the subject’s state of health after a temporary improvement.
2. Progression: Progression is defined as the process of developing or moving gradually towards a more advanced state. It basically means that the health of the subject under observation is improving.
3. Death: Death is defined as the destruction or permanent end of something. In our case, death will be our event of interest.

## REFERENCE

https://medium.com/@desouribac/source-unsplash-f917b1ce6673
https://www.cs.uct.ac.za/mit_notes/software/pdfs/Chp06.pdf