**SUMMARY**

**Purpose:**

In this assignment, you will create a Data Collection and Management Map. This is a high-level view of your entire data pipeline. Throughout the course you will be referring to this Map as you drill down into specific areas.

**Assignment Objective:**

Create a Data Collection and Management Map for the organization that you chose at the beginning of this course.

**Instructions:**

In this assignment, you are describing a business-related data problem that an organization is having. You will have received instructions from your professor about selecting your organization. Use the following prompts to frame your assignment. Please answer in paragraph format.

1. What is the name of your organization?

2. What business problem is your organization trying to solve?

3. Using the process map located in the text (Figure 2 Big Data Value Chain - Chapter 1) as a guide, please provide information on each of the following steps. If a particular step does not apply, please explain why.

a) Data Generation: What data do you need? Where are you going to collect the data from?

b) Data Collection & Transmission: What overall process will you use to collect the data? What format will the data come in? Are there any special considerations in transmitting the data to the storage center?

c) Data Pre-Processing: What pre-processing will you need to do on the data? Integration? Cleaning? Elimination of Redundant Data? Error Correction? Try to be as specific as possible.

d) Data Storage: What kind of storage system do you need? Which parts of ACID are required? How will you handle the CAP theorem?

e) Data Analysis: What kind of analysis do you plan to do with the data? What kind of tools might you need? What kind of visualization will you need?

f) Decision Making: How will decision makers want to receive the data?

# **CONTENTS**

[1. What is the Name of Your Organization? 1](#_Toc84000139)

[2. What Business Problem is Your Organization Trying to Solve? 2](#_Toc84000140)

[2.1 Background of the problem 2](#_Toc84000141)

[2.2. Problem definition 2](#_Toc84000142)

[2.3 What the company has done so far to solve its problem 2](#_Toc84000143)

[2.4 Conclusion 3](#_Toc84000144)

[3. Data Steps 3](#_Toc84000145)

[3.1 Data Generation 3](#_Toc84000146)

[3.3 Data Pre-Processing 5](#_Toc84000147)

[3.4 Data Storage 8](#_Toc84000148)

[3.5 Data Analysis 9](#_Toc84000149)

[3.6 Decision Making 10](#_Toc84000150)

[4. Reference 12](#_Toc84000151)



# **1. What is the Name of Your Organization?**



**Organization Name: NPhone**

* **NPhone** is an **international mobile company** headquartered in Dallas, USA. NPhone is one of the fast-growing manufacturers of portable batteries, hard disk, pen drives, smartphones and chipsets.
* In 2016, NPhone acquired D-Smart, a startup company **focused on smartphone design** **and development**. D-Smart offered innovative designs for recognized brands like Blueberry, Sunsang, Ni, Tomato to name a few. After successful merger of D-Smart and NPhone, NPhone launched its first model named as “Blade24” in USA and was great success in US market.
* NPhone was founded by real estate investors **Nicholas Benjamin** in **2010**.
* NPhone has **20% market share** for smartphone in **USA** and **5% in worldwide**.
* NPhone already has expanded to **12 countries** with recent attention on African markets.
* As per research of Counterpoint Research, India's smartphone market is growing at **14 per cent year-on-year** and set to hit a record high of **173 million units in 2021**. Hence NPhone wants to enter Indian market and explore the vast market. NPhone is well-known for its customer-oriented **designs and reliability**.
* The main reason for selecting a smartphone company is that, in India there are already a lot of major global players and hence this analysis is crucial for the **success** of the new company in India. It will be challenging to identify the **major factors and features** for the new smartphone model as the choice differ from region to region. Also, the availability of **vast dataset** from different source for the analysis.

# **2. What Business Problem is Your Organization Trying to Solve?**

## **2.1 Background of the problem**

NPhone is planning to expand their global presence and looking for new markets. They are planning to **enter Indian market** with their **new smartphone model**. NPhone wants to understand **customer demand and current market trend** of Indian market.

## **2.2. Problem definition**

* The design team is finding it difficult to decide on what **features and factors** affecting the customer **buying decision** in India and to understand customer needs **before designing a product**.
* Companies’ top management has targeted to achieve **10% of market** share in the first year itself.
* To study **socio-economic characteristics** which affects the **buying behavior** of the potential customers in India.
* Analyze the **feedbacks from** **customer** received for existing smartphones and **provide solutions** for the key problems.
* Company also wants to provide **Omni-channel retail experience** that focus on providing seamless customer experience whether the client is shopping online from a smartphone, a laptop/desktop or in a brick-and-mortar store.
* To formulate a **perfect product** for the Indian market in terms of features and provide **value for money**.
* Attract **maximum customers** with the **launch product**.
* Indian market is already dominated by several global players so it’s important to **understand the market**.
* Since this is launch product of the company, management wants **strong data support** to make decisions.
* To provide **best features and offers** for the launch product.

## **2.3 What the company has done so far to solve its problem**

* Company has started **investing in analytics** to understand the needs and demands of the people.
* Company has approached consultants to carryout a market study on Indian market.
* Feasibility study was carried out to decide on the entry into Indian market.
* As a marketing strategy, featured about “**NPhone Indian Plan 2022**” via Instagram, Facebook, Twitter and other digital magazines.
* **Vendor identification** has been started and **tied** up with a manufacturing unit in India.
* Earlier NPhone utilizes a team of designers to figure out what shoppers want to buy but now they started **investing on** **analytics** to understand the customer needs.
* **Collaborated with Flipkart,** one of the major online shopping portals in India with an agenda to **attract online shoppers** at the **introduction stage** of the brand.

## **2.4 Conclusion**

* NPhone wants to **penetrate into Indian smartphone market** with their launch product.
* **Attract maximum customer** at the introduction stage of the company in India.
* Identify key **features and factors** for the new model.

# **3. Data Steps**

To overcome the above listed concerns NPhone has to perform a bigdata value chain process.

## **3.1 Data Generation**

The big data value chain process **begins with data generation**. NPhone may generate relevant and needed data from the following information’s.

* Information of customers who showed interest in products via **social media**.
* Information of customers who have tried **login in NPhone website from Indian region**.
* **Sales record** of online and offline stores.
* **Previous online shopping records** of smartphones in India.
* **Analyze socio-economical** characteristics of smartphone buyers in India.
* **Customer feedbacks** for smartphone and key **customer concerns** in India.
* Analyzing **preferences of customers** across the globe by using big data analytics available Online.
* **Reviews, comments and ratings** on **online shopping portal**.
* Current smartphone **trends** and **developments** in India and across globe.
* **Purchase method and preference** of Indian smartphone customer i.e., Offline or Online.
* Attract new customers by providing **promotional offers and coupons** through online shopping portals and carryout social media campaigns.
* **Increase customer engagement** programs through referrals etc.
* Company can **host mega events** in Metro cities and partner up with TV Channels to increase brand reach.

All the collected data should be **reliable and accurate**, otherwise it will lead to **misleading insights and assumptions**. Data should be **examined and verified** before analyze.

**3.2 Data Collection & Transmission**

* This is the **second stage** of big data value chain process, here the data is obtained from all the possible sources.
* The data team collect and generate the data from all **reliable sources**.
* **Information of customers** who showed interest in products during social medias camping’s and during other customer interactions.
* Information of Indian customers from **company website**, while logging into the official online portal of NPhone the customers have to give their details such as name, email id contact number, gender etc.
* Information collected form **social media hashtag** data generation. Using basic detail of customers, company can identify the
* Type of customers
* Shopping pattern
* Major interests
* Changing fashion concepts of customers
* Big data analytics for understand the market trends by using **data-driven sentiment analysis** solutions on social media and other platforms will be helpful.
* Data from **comments, reviews and ratings** can be stored in a excel file
* **Web crawling** will help company to identify customers’ browsing patterns thus comprehending their choices.
* Use of tools like **Hadoop** will help the company to analyze the **pattern of customers** and use the insights to personalize the **offer to specific regions and customers**.
* Company’s **historical sales record**, **online shopping records, and financial records** can be used to analyze different demographic or gender to increase their popularity or sales.
* Designers need to focus on the **key features and kind of varieties** needed for the smartphone. They have a fixed set of resources like budget, time etc. Decisions should be backed up by data to decide how many models they need to launch and also need to forecast sales.
* **Online polls and survey** analysis can be used.
* Considering almost all the regions and vast online shopping, the flow of data would have a l**arge volume and high velocity**. Handling such a big data is not quite easy. So, this needs their own local storage space.
* In this process depending on the variety of data, it is collected in **different format**.

Considering the above information data can be gathered in below formats:

* Images (jpeg, png, jfif, dng, gif)
* Text, documents, scripts (Html, doc, pdf, bmp xml)
* Excel file (csv, xlsx)
* Videos (mp4, mpg, avi, mxf)
* Graphic images (TIFF, JPEG2000, PNG, BMP, GIF)
* Audio (WAVE, AIFF, MP3, MXF, FLAC)
* Database (XML, CSV, TAB)
* Once the data is collected, it is transferred to a data storage and processing infrastructure for further **processing and analysis**.
* While transferring the data the interim priority should be given to the **security** of the customer data. It should be alliance with the **privacy policy of the company** and it will be **end-to end encrypted** so that **security is highly** followed.

# **3.3 Data Pre-Processing**

* The data collected from various data sources may be **redundant, noisy and inconsistent**, hence, in this phase; the data is pre-processed to improve the data quality required for analysis.
* This also helps to improve the accuracy of the analysis and **reduce the storage expenses**.

The data can be pre-processed with the help of following steps:

**3.3.1 Integration**

This data pre-processing step involves in combining data collected from different sources and provide users with a standard format.

Some integration techniques are:

* Since the company is large, the data collected is uncontrollable by simple techniques. It needs **high end data integration tools** to integrate data from more discrepant source.
* The company can use a **data warehousing approach**, which **extracts, transforms, and loads** data from heterogeneous sources into a unique view schema so data from different sources become compatible.
* The data warehouse approach is **less feasible** for data sets that are frequently updated, requiring the **extract, transform, load (ETL)** process to be continuously re-executed for synchronization.
* The above techniques help company to integrate data **easily and efficiently**.



*Figure 3.3.1Simple Schematic Diagram of a Data Warehouse*

**Integration tools**

* On-premise data integration tool integrates data from the local sources and uses middleware software for connecting legacy databases.
* **Open-source data integration** tool is the best option in case you want to avoid expensive enterprise solutions. But using this tool will you have to handle the security and privacy of your data.
* Cloud-based data integration tool provides you ‘**integration platform as a service**’.

**3.3.2 Cleaning**

* Data cleaning is the process of preparing data for analysis to give the right results. There many ways to clean data depending on how it is stored and analysis followed.
* Information collected such as store records, online shopping records, social media records may contain many errors, missing columns, unwanted lines etc. unwanted recognition are cleaned in this section.

To clean up our data, we can follow some additional steps.

* **Eliminate Irrelevant Values**
  + There is always information that does not apply to all databases. Inactive data is the most likely does not match the context of the problem.
  + For example: once we have collected customer information in a retail store, an important fact we needed are; name, gender and age, the address or job title is not relevant to us. Deleting unwanted columns can reduce processing time for program, and give you an immediate output.
* **Remove Duplicates**
* Repeated values are the similar to irrelevant data. They only increase the volume of data, and increase the operating time of the system. But for larger datasets, it is not possible be find such datapoints. Sometimes when we combine multiple data sets or join them, recurrence is possible.
* Pandas' duplicate function () will report whether the given line is being repeated or not. All rows are marked as "false" to indicate that it is not a duplicate or "true" to indicate duplicate.
* **Outliers**
* As the data is collected from different countries there may occur outlier values, which may make complication in data. So, it can be eliminated.
* **Don’t Miss the Missing Values**
* There are always missing values in the database. If there is a specific column in the database has too many missing values, it would make sense to delete the entire column because there is not enough data to work with it.
* **Elimination of redundant data**
* Data redundancy occurs when the same piece of data is kept in multiple places or more separately places and it is common in many businesses.
* It makes data more complex so at this stage such information is deleted and the details can be merged into one sheet.

# **3.4 Data Storage**

* Data storage and management of data is a **continues process** which should satisfy the needs of applications or tools that require this data to be accessed at any time without delays.
* The data collected by NPhone, the volume will be **large and velocity** will be high. Hence the system chosen should provide reliable storage space and powerful access to the data. Therefore, NPhone can use cloud-based virtual data storage, so that all the data management team members can access the project easily and remotely.
* Setting up an online account with any of the **cloud services** is easy and cheap, also the data is highly secured.
* Cloud storage system is **convenient scalable and flexible** to use. It also provides an added advantage of recovery of data during a disaster.
* It is important to ensure that the database remains **consistent** before and after the transaction.
* To ensure the consistency of database, we will have the ACID (**ATOMICITY, CONSISTENCY, ISOLATION AND DURABILITY**) properties in our data transition, especially for the online sales:

1. **Atomicity**

* Atomicity makes sures that either the transaction occurs completely or it does not occur at all. “**all or nothing rule**” is the rule when comes to atomicity.
* Atomicity of the transaction is held responsible by Transaction Control Manager.

For example, if a sale happened is not recorded in inventory records there it forms a gap in

inventory.

1. **Consistency**

* This property ensures that integrity constraints are maintained and the responsibility lies with DBMS and application programmer.

For example, if a customer cancels an online order after confirming, it should be reflected in

customer account as well as in companies and besides the payment transaction can be tracked from by both parties involved.

1. **Isolation**

* Isolation makes sure that multiple transactions can occur simultaneously without

causing any inconsistency and the isolation responsibility lies with the control.

For example, if an online purchase is happened the details can be visible to all the branches.

1. **Durability**

* Durability ensures that all the modifications made by a transaction after its successful execution is written successfully to the risk and responsibility lies with the recovery manager to ensure durability in the database.

For example, if a customer accidently closed the online shopping portal after adding some

products in the card, the details will be automatically saved.

NPhone also is planning to follow the **CAP Theorem** (Consistency, Availability, Partition Tolerance).

* In normal operations, your data store provides all three functions. But **network failure is not an avoidable scenario** for any organization which forces them to choose one between Consistency and Availability.
* NPhone has chosen **consistency over availability** all users need a consistent view of the data in their application more than availability.
* These systems are not completely available but **strongly consistent** which is the company’s priority.

# **3.5 Data Analysis**

The company needs solutions to its problems. To solve these problems, a comprehensive data analysis will be done. Such as:

**Community Analysis**: To assess most rated items and locations in various areas

**Trending Product Analysis**: To figure out which features are requested most by the customers with respect to region.

**Social Media Analysis**: To collect client’s data who have looked through the online media records of the organization.

**Behavioral Analysis**: To comprehends the purchasing conduct, design, intention of the customers.

**Online Sale Analysis**: To see how quick the interest for online deals is expanding. Also, to comprehend the most moving class of features.

**Precaution Analysis**: To decide constantly what sort of additional actions can be taken to ensure the staff, clients and general wellbeing.

For data analysis, the company can use below mentioned analysis tools:

**Tableau**: will be used for visualization that is to generate best visualization using Graphs, charts, maps, and diagrams.

**SQL**: will be used for adding queries to and deleting queries from the dataset together with joining the tables.

**Excel**: With the help of excel customer data’s can be easily stored and which can be used for various purpose.

**R**: Statistical analysis will be carried out with the help of R. For instance, variance, mean, standard deviation, etc. Plots like boxplot can be generated for the better understanding.

**Oracle**: will be used for classification, prediction, regression, and specialized analytics to generate insights, make better predictions, target best customers, identify cross-selling opportunities, and detect fraud.

# **3.6 Decision Making**

* Based on the analysis and the **visualized results**, the top authorities of the company can decide the perfect trade off which will give them market advantage.
* The details of a particular problem can be analyzed to understand the causes of the problems take informed **decisions and plan** for necessary action.
* And to do that, the analysis team needs to use various types of tools. Such as Tableau, Excel, etc. Top management want to receive the information in the form of Diagram.
  + Shopping pattern of customer’s diagram.
  + Most and least preferred features of products
  + Profit margin from every region.
  + Pie charts, sales line graphs, histograms.
  + Inventory records.

The decision-makers want to receive the data in an informative way such as:

* Histogram of most moving category of product.
* Low and high profitable cities and branches
* Excel data of the selling stores and their contribution.
* Line graphs of the sales Vs. season.
* Sales Map representation
* Final dataset ER diagram

# **4. Reference**

* https://academic.oup.com/comjnl/article/4/2/103/383168
* <https://www.coursera.org/learn/introduction-to-data-analytics>
* <https://www.astera.com/type/blog/understanding-data-mapping-and-its-techniques/>
* <https://www.talend.com/resources/data-mapping/>
* <https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/>