# **Problem Statement**

Knapsack is an e-commerce startup that sells backpacks. It has a website Knapsack.com and a mobile app(ios and android) Knapsack. It has been incredibly successful and is growing at a fast pace. You are a data product manager and you have to create a Data Strategy proposal for the next version of Knapsack. Bring it all together from the previous exercises and complete the template for Knapsack Data Strategy Proposal and fill in gaps.

Here are few things to consider while writing this proposal:

1. The Business Intelligence solution is a proprietary one with limited functionality.
2. There is a lot of talk about event data. Leadership is still not clear what all data might be needed and how it can be used
3. Knapsack is soon going to be launched in European Markets.
4. Leadership is concerned about security and compliance, can you recognize main categories of PII and PCI data.
5. Leadership is weighing upon the option of moving to cloud for scale but hasn’t been able to decide either way.
6. A recent article on a site mentioned competitors' efforts of building a backpack that we need in the age of IoT. How does this impact some of our strategy decisions?
7. Leadership is concerned about scaling if they still keep all pieces of Smart Knapsack in house; especially product design and development, sales and marketing, customer support and manufacturing. All these teams are accumulating tons of data from their day to day operations and are collecting all of the event data from our connected knapsacks.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Knapsack Data Strategy Proposal**

## **Introduction**

Our overwhelming success at Knapsack has made it imperative that we reconsider some of our data infrastructure decisions so that we do not get overwhelmed with our data needs. As Data Product Manager of Knapsack, I took on the challenge and responsibility of exploring every nook and corner of the Data Infrastructure needs. Here I present to you my findings along with my suggestions that will help us scale not only for our current needs but will also support our future growth.

## **Section 1:** Knapsack Data Pipeline

Our Knapsack Data Pipeline looks like following:

[Paste picture of illustration that you created]

Knapsack Data Pipeline components can be categorized in following buckets:

[Fill and complete following list]

Data Origin

* Website
* Mobile app
* IoT (data taken by the backpack)
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Data Processing

* Ingestion
* \_\_\_\_\_\_\_\_\_\_\_\_\_
* \_\_\_\_\_\_\_\_\_\_\_\_\_

Data Storage

* Data Lake (?)
* Data warehouse

Data Destination

* Machine Learning Application
* Business Intelligence
* \_\_\_\_\_\_\_\_\_\_\_\_

## **Section 2:** Data Consumers

**Data Consumers**

Our primary business goals, data consumers and data use-cases are as following:

[Fill in following table]

|  |  |  |
| --- | --- | --- |
| **Business Goals** | **Data Consumers** | **Data Use case** |
| Having a working product | Engineering | Monitoring site and app performance. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Data Needs**

Our primary data needs can be encapsulated as:

[Fill in the following table]

|  |  |  |
| --- | --- | --- |
| **Stakeholder** | **Data Type Needed** | **Data Elements Needed** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Inventory Management Team | Event + Entity Data | Event Data: Event ID, Time Stamp, Event Type, Batch Number  Entity Data: Time stamp when order was placed, time stamp when order was dispatched form manufacturer facility, time stamp when order was delivered to warehouse, batch number |

**Data Model**

Relational data model that we need for marketing, finance and customer care use cases is:

[Fill the following tables and necessary information to complete this section]

Customer

Primary Key: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | Email |

Customer Demographics

Primary Key: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Gender |  |  |

Product

Primary Key: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Product ID |  |  |  |  |  |

Order Details

Primary Key: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Foreign Key: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Foreign Key: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Product ID |  |  |  |  |

## **Section 3:** Data Producers

**Data Producers**

[Is anything missing from the list? If yes, find out what and complete the list]

Our data producers are:

1. Website: Knapsack.com
2. Saas Solutions
   1. Simply Accounting
   2. Just Sale
3. Storage Devices
   1. Relational DB
   2. Data Lake

**Entity Data Producers**

[Is anything missing from the list? If yes, find out what and complete the list]

Entity Data Producers are:

1. Website and Mobile app: All our transactional data is entity data.
   1. Customer registering and creating account
   2. Customer buying backpacks
2. Relational Database storing transactions that will provide us following insights:
   1. What was bought ?
   2. By whom ?
   3. When ?
   4. How much ?

**Event Data Producers**

[Is anything missing from the list? If yes, find out what and complete the list]

Event Data Producers are:

1. Website and Mobile app: Data generated by customer interaction.
   1. Users searching backpack
   2. Clicking through product pictures
   3. Navigating through various sections and pages
2. Connected Backpacks: Our connected backpacks are IoT devices and have components that generate event data.
   1. Anti-theft device
   2. Location Tracker
   3. 911 Button

**Backend Data Producers**

Often left out or forgotten are backend data producers. We have few and would need to invest in some.

***Have***

[Is anything missing from the list? If yes, find out what and complete the list]

1. OLTP: We have OLTP system and storage.
   1. It keeps our transactional data ACID compliant
   2. It has a relational DB
2. CRM: Just Sale helps us with our marketing and sales initiatives. Of the many things it does here are few:
   1. Monitors our marketing channels
   2. Tracks our marketing campaigns

***Need***

[Build vs Buy Decisioning]

Our current business intelligence application is a small off the shelf solution with limited capabilities. Some of the problems we are facing with it are: [List down reasons why buying might not have worked]

1. It cannot be customized to solve all our use cases.
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

I ***propose*** we build OLAP functionality in house to support our BI and to build a stronger analytical muscle. It will provide us with following benefits: [List down reasons why building might be better]

1. Control over cost
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Event Data Needs**

We want to track following KPIs:

1. Session Length on website and app to measure how engaging and relevant content is.
2. Click Through Rates on the website and app for new launched products to gauge interest of customers in new products.
3. Conversion Rate around how many backpacks are we able to sell.
4. Daily Active Users on app to monitor our growth
5. Bounce Rate on website, to gauge customer engagement

For doing so we will need following event data:

[Fill and complete following list]

Session Length

* Events for when customer gets to the site
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Click Through Rate

* Click event on the new product pic on site
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Conversion Rate

* Event for customers getting on the site
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Daily Active Users

* Event for when customers launch app with unique user id

Bounce Rate

* Events for when customer gets to the site
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Types of Data**

*(@Kanpsack)*

Before recommending what all pieces of technology we need I would like to elaborate about type of data that we are generating in Knapsack [Fill and complete following list]

1. Structured
   1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. Semi-structured
   1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. Unstructured
   1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **Section 4:** Data Processing and Data Storage

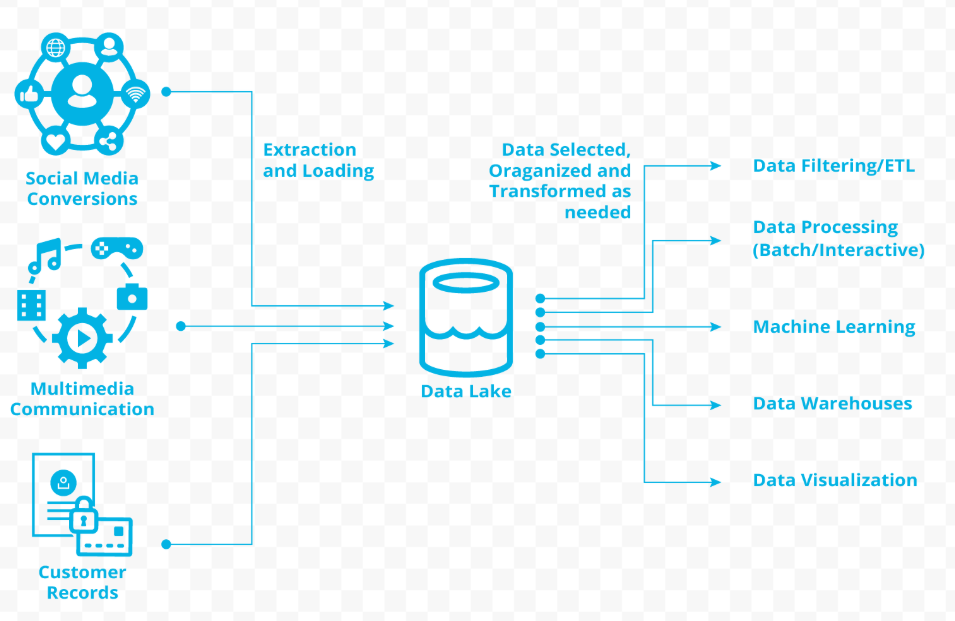
**ETL and Data Warehouse**

To implement our OLAP Data Pipeline we need ETL processing since data we need to use is structured and batch in nature. Our use cases for OLAP need batch processing hence ETL works best. Here is our OLAP Data Pipeline.

[Paste picture of illustration that you created]

**ELT and Data Lake**

We are collecting a lot of unstructured and semi-structured data that we store in Data Lake. It is critical for the success of our data infrastructure that we apply appropriate processing strategy for this piece. I ***propose*** that we reconsider the traditional architecture and adopt an ELT pipeline. Here is what the the pipeline can look like:



We can collect all our social media chatter , connected devices(our smart knapsack) data and also the data from customer tables. All the data from different sources will be extracted and then loaded to the data lake. Data from data lake can be fed to all of our other applications. We can build an ETL pipeline on top of Data lake instead of fetching data from different databases. This will save on time and infrastructure costs. A centralized data lake will provide us following advantages: [List down advantages of using ELT pipeline with Data Lake]

1. Easy integration of new data sources as data lakes do not have strict data format requirements
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Meanwhile we will have to be careful because of the following reasons: [List down concerns for using ELT pipeline with Data Lake]

1. Our data lake should not become a data graveyard, where all data is dumped forever without anybody needing it.
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **Section 5:** Data Infrastructure and Strategy

**Security and Compliance**

We are primarily an e-commerce company. Our control and security requirements are not as stringent as of financial institutions or of defence companies. That does not mean we do not have sensitive data. We have PII data and data that falls under PCI compliance.

[Complete the following table]

|  |  |  |
| --- | --- | --- |
| Data | PII or PCI | What we need to do? |
| Customer Personal Information | PII | Data storage should be GDPR and CCPA compliant |
|  | PCI |  |
|  |  |  |
| Supplier Payment Information | PCI | Appropriate controls in place to follow PCI DSS core objectives |
| Employee Personal Information |  |  |
|  | PCI |  |

**Cloud or On-Prem**

On-prem gives us full control over how we design and maintain our tech infrastructure. This comes at a cost. We have to buy physical infrastructure pieces and hire an army of professionals to keep them up and running. If we need to scale down in the off season we are wasting our resources and in case we need to exponentially scale up in peak season we might not be able to do so.

On the other hand cloud has following advantages: [List down advantages of moving to cloud]

1. We can scale up and down as needed, and pay only for the services/infrastructure used.
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

I ***propose*** we should move to the cloud.

**Smart Knapsack and Scalability**

Smart Knapsack has been the most successful product from our catalogue. This speaks about the various teams that came together to make it a fruitful venture. We have threats in the environment, to grow and scale we need to evaluate a few things.

What should be our focus area?

We are an e-commerce business, like many others. [What is our core functionality] \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Should all teams be collecting all the event data?

This is concerning. While we have had Data Lake, we never used it efficiently. [Why idea of every team collecting data is wrong, what kind of classic data problems it might cause]\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

To scale and create the next magical product without burdening our Data Infrastructure I ***propose*** following: [Apply learnings around insource and outsource to complete following list]

1. Focus on core functionality. And out-source some non strategic functions. To start we can outsource following:
   1. \_\_\_\_\_\_\_\_\_\_\_\_\_
   2. \_\_\_\_\_\_\_\_\_\_\_\_\_

This will give provide us multiple benefits such as:

1. Immediate scale to address concerns of growing customer base.
2. Opportunity to scale our manufacturing and fulfill demand.
3. \_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_
5. \_\_\_\_\_\_\_\_\_\_\_\_\_

We should not out-source:

1. \_\_\_\_\_\_\_\_\_\_\_\_\_
2. \_\_\_\_\_\_\_\_\_\_\_\_\_

The reasons why we should not out-source the above

1. We want to keep our Smart Knapsack IP inhouse
2. Sales and marketing needs a good know-how of what we are as a company, what we represent. The more assimilated they are in company culture the better they represent us.
3. \_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_
5. \_\_\_\_\_\_\_\_\_\_\_\_\_

2. I cannot re-emphasize enough why we need to move to the cloud. It can help with

1. Launch in European market: \_\_\_\_\_\_\_\_\_\_\_\_\_
2. Security and compliance: \_\_\_\_\_\_\_\_\_\_\_\_\_
3. Competitor threat: \_\_\_\_\_\_\_\_\_\_\_\_\_
4. Scaling: \_\_\_\_\_\_\_\_\_\_\_\_\_

## **Summary**

In our journey of building Knapsack as a brand and company we made a lot of good decisions, but what got us here might not take us to the future. To make it bigger and better we have to decide on some criticat data strategy pieces. This proposal is an attempt to bring all those pieces together with the right context to empower our leadership to march us forward.