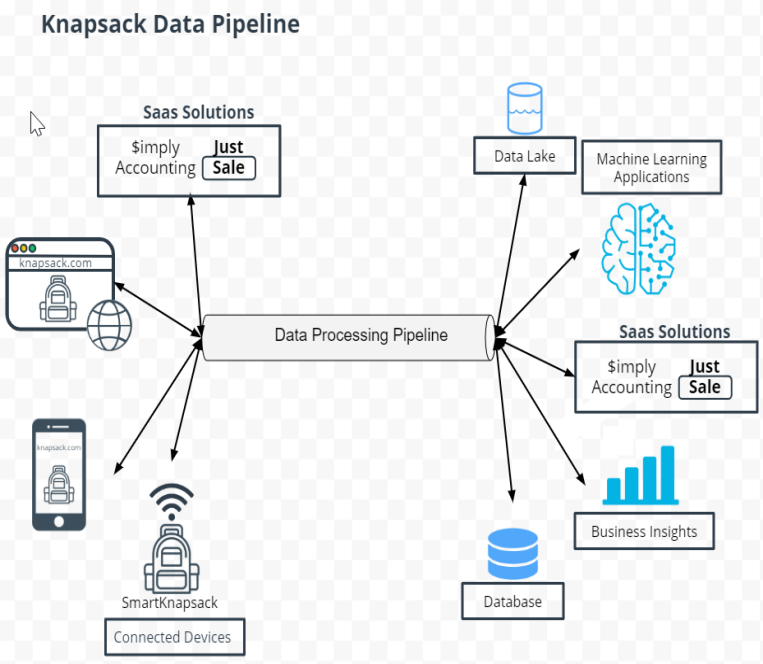
Knapsack Data Strategy Proposal

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:



Knapsack Data Pipeline components can be categorized in following buckets:

Data Origin

- ❑ Website
- ❑ App
- ❑ SaaS Solutions
- ❑ Connected Backpacks

Data Processing

- ❑ Ingestion
- ❑ Transformation
- ❑ Data Delivery

Data Storage

- ❑ Data Lake
- ❑ Database

Data Destination

- ❑ Machine Learning Application
- ❑ Saas Solutions
- ❑ Business Intelligence Application

Section 2: Data Consumers

Data Consumers

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

Business Goals	**Data Consumers**	**Data Use case**
Having a working product	Engineering	Monitoring site and app performance.
Improving product	Product Management	Identifying customer pain-points
Getting new customers and retaining old	Marketing	Targeted advertising
Predicting P&L	Finance	Monitoring current P&L
Smooth Product Delivery	Logistics Team	Identify delivery Bottlenecks
Sufficient Inventory	Inventory Management Team	Monitor inventory Lead Time
Addressing customer grievances	Customer Care	Provide personalized responses to the customer

Data Needs

Our primary data needs can be encapsulated as:

Stakeholder	**Data Type Needed**	**Data Elements Needed**
Engineering	Event Data	Event ID, Timestamp, Event type
Product Management	Event Data	Event ID, Time stamp, Event Type, Event Page
Marketing	Entity Data	Customer name, email, phone, address, customer order history
Finance	Entity Data	Aggregated Transactional Data, Number of transactions, cost of transactions, retail price, taxes, other charges if any.
Logistics Team	Event + Entity Data	Event Data: Event ID, Time Stamp, Event Type, Parcel Number, Hop Number Entity Data:Parcel Number, Hop Number, Time at Hop.

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:

Customer

Primary Key: Customer ID

Customer ID	First Name	Last Name	Address	Email

Customer Demographics

Primary Key: Customer ID

Customer ID	Age	Gender	Marital Status	Parental Status

Product

Primary Key: Product ID

Product ID	Type	Color	Size	Material	Cost Price

Order Details

Primary Key: Order ID

Foreign Key: Product ID

Foreign Key: Customer ID

Order ID	Date	Product ID	Customer ID	Quantity	Retail Price	Tax Rate

Section 3: Data Producers

Data Producers

Our data producers are:

1. Website: Knapsack.com
2. Mobile App: Knapsack
3. Saas Solutions
    a. Simply Accounting
    b. Just Sale
4. Connected Backpacks
5. Storage Devices
    a. Relational DB
    b. Data Lake

Entity Data Producers

Entity Data Producers are:

1. Website and Mobile app: All our transactional data is entity data.
    a. Customer registering and creating account
    b. Customer buying backpacks
2. Saas: Simply Accounting and Just Sale
    a. Simply accounting will have tables around all the monthly financial transactions of Knapsack
    b. Just Sale will have data around marketing channels and associated campaigns running on them
3. Relational Database storing transactions that will provide us following insights:
    a. What was bought ?

b. By whom ?
c. When ?
d. How much ?

Event Data Producers

Event Data Producers are:

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

1. OLTP: We have OLTP system and storage.
    a. It keeps our transactional data ACID compliant
    b. It has a relational DB
2. ERP: Simply Accounting is our ERP. Some of the things it does are:
    a. It creates reports of transactional data
    b. It automatically makes balance sheet entries
3. CRM: Just Sale helps us with our marketing and sales initiatives. Of the many things it

does here are few:

a. Monitors our marketing channels
b. 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:

1. It cannot be customized to solve all our use cases.
2. To get all the required functionalities we have to pay a hefty amount, total cost of ownership is not favorable.
3. Deep understanding of our business needs is required to provide a good analytical solution.
4. More and more data consumers need to look at data across multiple dimensions, in our current solution we need to pay per user. We are already at the limit.
5. For integrating every new data source we have to pay.

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:

1. Control over cost
2. Fully customized to our needs
3. Long term this sits at the core of how we take business decisions, if we have inhouse expertise to build state of the art analytics capabilities we can make better decisions.
4. We can serve all our internal data consumers.
5. We will be able to integrate and grow as per our need, we do not need to depend on our vendor for every new data integration.

Event Data Needs

We want to track following KPIs:

1. Session Length on website and app
2. Click Through Rates on website and app for new launched products
3. Conversion Rate around how many backpacks are we able to sell.
4. Daily Active Users on app
5. Bounce Rate on website

For doing so we will need following event data:

Session Length

- ❏ Events for when customer gets to the site
- ❏ Event when customer closes the site
- ❏ Event when app was launched
- ❏ Event when user becomes inactive on app
- ❏ Timestamp for all such events

Click Through Rate

- ❏ Click event on the new product pic on site
- ❏ Click event on the new product pic on app
- ❏ Event for when customer gets to the page on which new product banner is

Conversion Rate

- ❏ Event for customers getting on the site
- ❏ Event for customers launching app
- ❏ Event when customer presses buy button

Daily Active Users

- ❏ Event for when customers launch app with unique user id

Bounce Rate

- ❏ Events for when customer gets to the site
- ❏ Event when customer closes the site along with page info
- ❏ Event when app was launched
- ❏ Event when user becomes inactive on app along with page info

Types of Data

(@Kanpsack)

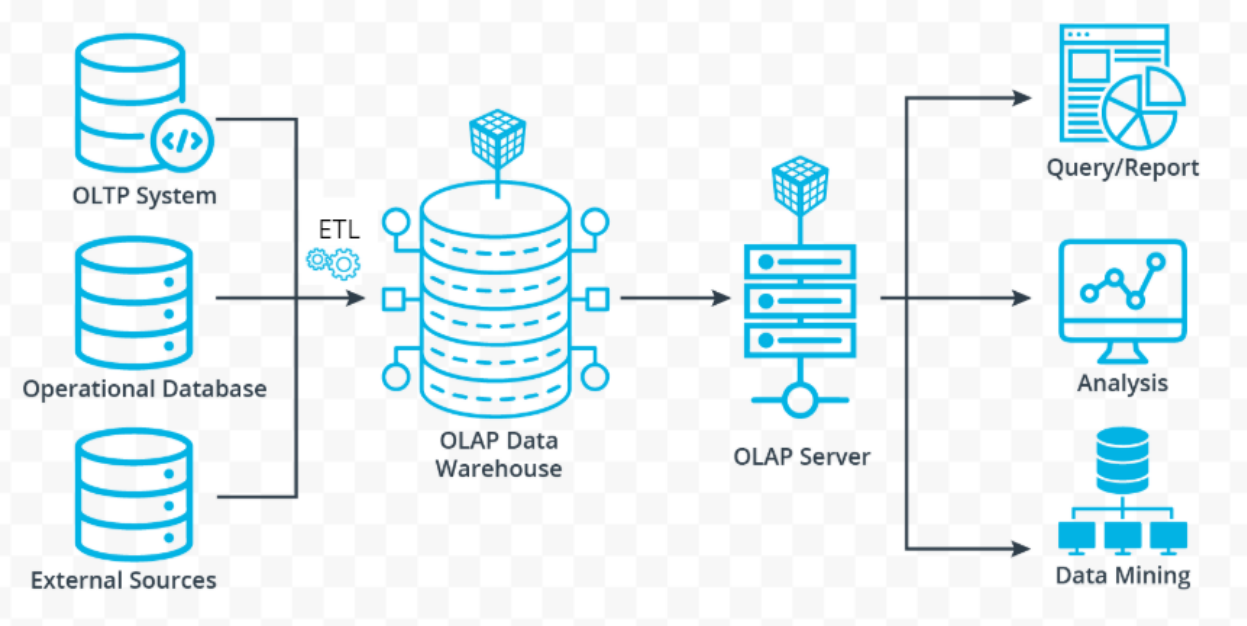
Before recommending what all pieces of technology we need I would like to elaborate about type of data that we are generating in Knapsack

1. Structured
    a. Entity and event data -website and app
    b. Event data - Connected Knapsacks
    c. Relational DB, ERP and CRM
2. Semi-structured
    a. Site and app communicates to servers through JSON and XML
    b. Customer emails stored in Data Lakes for sentiment analysis
3. Unstructured
    a. Customer Reviews, Product Pictures, Videos, Product Mentions on social media platform
    b. Some other data stored in Data Lakes

Section 4: Data Processing and Data Storage

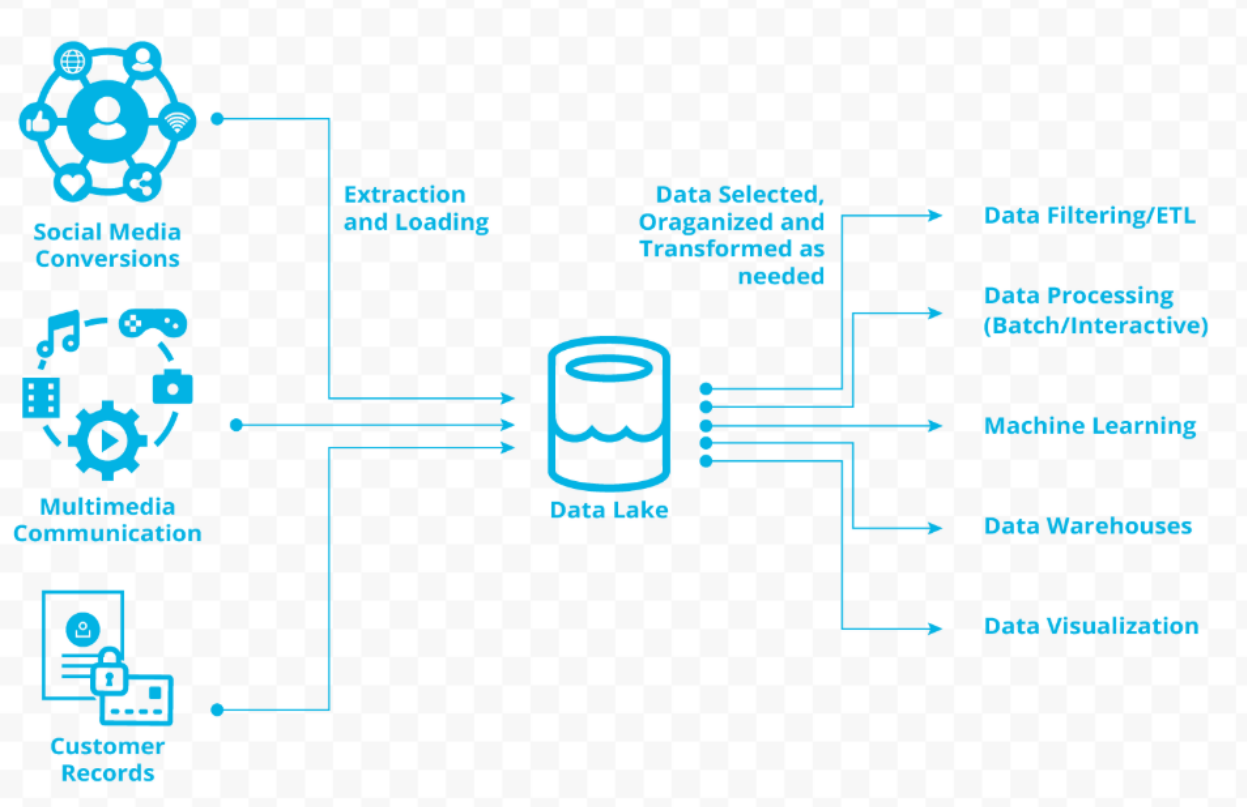
ETL and Data Warehouse

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.



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:

1. Easy integration of new data sources as data lakes do not have strict data format requirements
2. Less time to integrate new data sources
3. Data in data lakes will be in its natural form that will help us use and process data in multiple ways (based on how applications need it).
4. It will help us scale, as our data is growing day by day, we need a scalable solution.

Meanwhile we will have to be careful because of the following reasons:

1. Our data lake should not become a data graveyard, where all data is dumped forever without anybody needing it.
2. No data format requirements does not mean basic data hygiene is not followed, because even with best of the technologies the principle of garbage in garbage out still stands.
3. Strict data governance and monitoring will be needed
4. Data compliance might be challenging

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.

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

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:

1. We can scale up and down as needed, and pay only for the services/infrastructure used.
2. Our teams can focus on innovation and core business functionality instead of focusing on making underlying infrastructure work.
3. Cloud technologies have evolved tremendously and give us an opportunity to use the latest and greatest.
4. As we are planning to expand to other markets, cloud becomes even more relevant. We do not need to worry where data centers should be built, depending on local laws data can be saved in a particular region if need be.

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 gives us an edge is the unique range of products that we have, especially our in-house innovation Smart Knapsack. It is our point of differentiation and unique value proposition. Our focus area should be product design and innovation.

Should all teams be collecting all the event data?
This is concerning. While we have had Data Lake, we never used it efficiently. They have built their own eco-systems around data, this has created silos and duplication. All the teams should not be collecting all the event data and should not build their data infrastructure in isolation. This is a waste of resources and will cost us.

To scale and create the next magical product without burdening our Data Infrastructure I ***propose*** following:

1. Focus on core functionality. And out-source some non strategic functions. To start we can outsource following:
    a. Customer Support
    b. Manufacturing

    This will give provide us multiple benefits such as:

a. Immediate scale to address concerns of growing customer base.
b. Opportunity to scale our manufacturing and fulfill demand.
c. We can focus on our core functionalities and innovate.
d. We might even be able to save on cost if we tap right manufacturers who have already established economies of scale.
e. We can tap into 24x7 support centers established across the globe that will help us in taking care of our new European customers without burdening our US support centers.

We should not out-source:

a. Design and development of new Knapsack models
b. Sales and marketing

The reasons why we should not out-source the above

a. We want to keep our Smart Knapsack IP inhouse
b. 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.
c. These are core to our business.
d. Over time we have built a strong team of smart individuals who have already given us a headstart in the market. We just need to focus on their growth and development, smart people attract smart job aspirants. Eventually we could build a talent pool that in itself may prove to be our differentiation in the market.

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

a. Launch in European market: We can simply deploy an instance and do not need to procure hardware in Europe in order to launch our website and app there.
b. Security and compliance: Cloud service providers have an alacarta of time tested services and tools that can be readily used. Cloud will also help with data privacy and data residency laws. (What data can cross borders and what cannot?)
c. Competitor threat: We can focus and innovate more and faster, maintaining our first mover advantage in the world of smart backpacks.
d. Scaling: Need based scaling is the biggest advantage of moving to cloud.

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.