Introduction to Data Pipelines Lesson

# Basic Data Pipeline Exercise

Create a basic data pipeline graphic for what Knapsack's data pipeline might look like. It should include more than 3 parts. \*This is more of a pre-quiz to see what you know at the start. It's ok if you are not sure of your answer.\*

You should also create an outline of the pipeline below. You can use any drawing tool you would like and Google Drawing, MS Paint, Visio ETC…

You will not be able to paste your image here. Instead, compare the image you created with the solution on the next page.

# Knapsack's Data Pipeline Expanded

At Knapsack, we decided to focus on the core business of selling backpacks. Hence we decided to use cloud-based third-party solutions for some business units.

We use “SimplyAccounting” for accounting and “JustSale” for managing marketing campaigns.

Also, we have recently launched a beta version of our “Smart Knapsack,” that is in the market with few beta customers, and we are continuously monitoring its performance. The Smart Knapsack has the following features:

- Anti-theft device

- Wifi signal booster

- Solar panel to charge devices

- Retractable light

- One button 911 dial with location tracking.

Try to recreate the Knapsack data pipeline. Can you identify SaaS and IoT components? What do you think could be some potential data problems for Knapsack?

At Knapsack, we decided to focus on the core business of selling backpacks. Hence we decided to use cloud-based third-party solutions for some business units.

# Advanced Data Pipeline Exercise

We use “SimplyAccounting” for accounting and “JustSale” for managing marketing campaigns. Also, we have recently launched a beta version of our “Smart Knapsack,” that is in the market with few beta customers, and we are continuously monitoring its performance. The Smart Knapsack has the following features:

- Anti-theft device

- Wifi signal booster

- Solar panel to charge devices

- Retractable light

- One button 911 dial with location tracking.

Try to recreate the Advanced with the new components Knapsack data pipeline around storage. Can you identify potential data storage needs? What do you think could be some potential data problems for Knapsack?

# Alex Headphones Exercise

You are the product manager of Alex Headphones. These are new connected headphones that allow the user to connect to any device or use the headphones over a wireless network and can connect and use any voice assistant. The can provide the user with just in time updates to their daily schedule and connect to any music service too. Setup and control happens via either a native iOS/Android App or via a web interface. Also your company has chosen to use AWS and its products for its data pipeline.

*If you are more familiar with other similar products you can use those in place of AWS. However, the solution is given using AWS*

You are tasked with creating the outline of a data pipeline with all its components. Along with your imagination, use what you have learned in this lesson and create the model. Also reflect on why these components are needed. You may also want to create your outline visually as we have done with the exercises previously. You can use a drawing tool like draw.io or Google drawing or create a simple text outline to complete this activity.

Your outline should include:

* Data Origin Components
* Data Destination Components
* Data Processing Components
* Data Storage Components
* SaaS if any?
* Connected devices if any?

# 

Data Consumers

# ID data consumers

Remember, Knapsack is a start-up. We have limited resources; our first few goals are:

1. Having a working product

2. Getting new customers and retaining old

3. Predicting P&L (Profit & Loss) growth

4. Smooth product delivery

5. Having sufficient inventory

6. Addressing customer grievances

7. Improving product

Can you identify one of our data consumers for each goal? You can do this in the space below or you can graphically depict it using a tool like Google Drawing, MS Visio, Sketchpad or Lucidchart.

# Identify data Use-case

We know Knapsack’s business goals, we know Knapsack’s data stakeholders.

1. Having a working product: Engineering Team

2. Getting new customers and retaining old: Marketing Team

3. Predicting P&L growth: Finance Team

4. Product Delivery: Logistics Team

5. Having sufficient inventory: Inventory Management Team

6. Addressing customer grievances: Customer Care Team

7. Improving product: Product Management Team

What are the data use cases for these data consumers? Which category does each of these use cases fall into?

# Entity Data Quiz

Please use the **Entity Data Quiz Tab** in the **Data Strategy Class Exercises spreadsheet** to help you complete this activity.

For a product of your choice, create entity model attribute tables for both products and orders. Use your imagination for what attributes could be associated with a product and an order. Use the areas below each table type to create a table for the different attributes for your product and customer. Also remember to include which columns are Primary and Foreign Keys.

**HINT: You can use the customer table discussed previously as a reference point.**

# Recognizing Data Needs Exercise

Use the table in the **Data Strategy Class Exercises** spreadsheet in the **Recognizing Data Needs Quiz tab** to complete this activity.

For all the Business Goal, Stakeholder, and Use Cases identified in the table, determine

1. The kind of data that will be needed(entity/event).
2. Now that you have recognized the type of data, identify what minimum set of data elements are needed in each case.
3. Choose one of the following, Marketing, Finance, or Customer Care use cases, and create entity data model tables with at least two tables and four attributes per table. Make sure to identify the Primary and Foreign key(s) for each table.

**Standout Suggestion**: Use a diagram tool like draw.io or Google Drawing to map this out visually.

# Car Sharing Exercise:

**Scenario**

You have just been hired at a new start-up looking to disrupt the car rental industry. This company offers peer-to-peer car-sharing where a person who has a car can rent it out to someone who needs a car for a specified amount of time. In this situation, both the car owners and renters are customers.

**Additional Details to Consider**

Renters searching for cars might want a certain type or model in a certain location for a certain amount of time.

The owners might want to add unique details about their rentals and get information about the renter before agreeing to rent out their car. There are many other considerations like verifying customer's driving records, insurance, and car details as well, but you do not need to get too detailed for this exercise.

# Consider the car-sharing company from above and Identify the following:

# List at least 4 data needs for this company

# Identify at least three data consumers/stakeholders

# Identify at least one data use-case of each data stakeholder

# Determine at least 3 data elements needed for each use-case

# Identify if these data elements are entity-based or event-based (at least one of each)

# Complete an entity data model table for at least 3 entity data elements

# 

Data Producers

# Knapsack Event Data Quiz 3

Create an event model table that aggregates the data for each of the following data types:

1. Active users

2. Most popular feature

3. Least popular feature

4. Daily active users

5. Session length

If you can determine the totals for each of the above, what is the answer? If you’re unable to determine the totals, why are you not able to do so?

**Please use the data from the tab Knapsack Event Data to complete your data analysis of Knapsack's Event data. Located in the Tab Knapsack Event Data Quiz provided in template spreadsheet in the classroom resources to complete this exercise.**

# Knapsack Data Producers

For Knapsack Data Producers identify following:

* Entity Data Producers?
* Event Data Producers?
* OLTP needed?
* OLAP needed?
* ERP needed?
* CRM needed?

# Knapsack Data Continued

## Knapsack

Knapsack has the following data producer: website, app, connected devices, ERP, CRM, relational Database and a data lake.

Answer the following:

1. Which types of data might these producers create is entity and which is event?

2. Which category would they fall into: Structured, Semi-structured, Unstructured?

Data Strategy

## Choose Your Processing Elements

At Knapsack we are building an OLAP ecosystem.What that means is we will have a Data WareHouse and an OLAP server on top of it. Knowing what we know so far, what are all of the sources of data that should feed into this OLAP.

1. At Knapsack we are building an OLAP ecosystem.What that means is we will have a Data WareHouse and an OLAP server on top of it. It will be used for business reporting and understanding historical sales trends. Knowing what we know so far what all sources of data should feed into this OLAP.

2. What type of processing is required batch processing or stream processing for the above use case? Why?

3. Do we need an ETL or ELT pipeline? Why?

4. For the above use case can you create an illustration of a data pipeline which has following:

* Data Producers
* Processing elements
* Data Consumer: OLAP Data Warehouse

## Knapsack Data Storage Quiz

Now it's your turn to figure out what Knapsack's Data storage needs might look like here's what we know so far:

Knapsack's Entity Data

For Knapsack pick a DB for each of the following:

* Structured Data
* Semi-structured Data
* Unstructured Data

# Knapsack Data Strategy Proposal