1. Understanding the business - End to End AWS Glue project

Introduction

For implementing this project, I have created a dummy company called Pure Fresh. They are a milk sourcing and distribution company. For your reference, you can think of it as a smaller version of AMUL if you are from India. The reason I created this business was that I wanted to mock daily data generation and spice up the project. Instead of using some random datasets from the internet, I decided to create it myself. The data is being generated using Python. 🐍

I am aware of this industry, and the data is synced properly. So it's not random data generation; it's well-fabricated. I have spent considerable time creating this business. The mock data is very well created, and I have also added some impurities to it. This process gave me exposure to BA-related work, and for data generation, I have created APIs, which was a good learning opportunity as well. I have learned and implemented a lot in this project. Although I have worked on real projects, this one was where I did everything from scratch. With that, I will step back and let the flow begin. From now onwards, I will be playing the role of a data consultant, and our client is Pure Fresh. 😊

What is Pure Fresh?

Let's look at their company profile.

Overview:

Pure Fresh is a small but rapidly growing dairy company dedicated to providing high-quality dairy products sourced directly from local farmers. With a commitment to freshness and purity, Pure Fresh operates a streamlined process from collection to distribution, ensuring that every product meets the highest standards. 🥛

Mission:

Our mission at Pure Fresh is to deliver the freshest and most nutritious dairy products to consumers while supporting local farmers and communities. As we grow, we aim to expand our reach while maintaining our core values. 🌱

Our Process:

1. Collection from Farmers:
    - Pure Fresh operates a few collection centers strategically located near dairy farms.
    - Specially equipped vehicles collect fresh milk from farmers, ensuring a seamless and efficient process.
2. Chilling Center:
    - Upon collection, the milk is swiftly transported to our modest but efficient chilling center.
    - Here, the milk is rapidly chilled to preserve its freshness and quality. ❄️
3. Production at the Factory:
    - At our small yet advanced production facility, the chilled milk undergoes meticulous processing.
    - Pure Fresh currently offers a select range of dairy products, including:
        - Fresh Milk
        - Skimmed Milk
        - Flavored Milk
        - Cottage Cheese
        - Yogurt

Quality Control:

- Pure Fresh maintains rigorous quality control measures at every stage of production.
- Our dedicated team of experts ensures that only the finest ingredients are used, resulting in products that meet and exceed industry standards. ✅

Logistics and Distribution:

- The company manages its logistics efficiently, ensuring timely delivery from the factory to local distributors.
- Pure Fresh is working to establish stronger partnerships with reliable distributors to increase the availability of our products. 🚚

Our Team:

- Our small team at Pure Fresh is comprised of skilled professionals dedicated to excellence.
- From farm collection to factory production and distribution, each employee plays a crucial role in delivering the best to our customers. 👩‍🌾👨‍🔬

Customer Satisfaction:

- Pure Fresh values customer satisfaction above all else.
- We actively seek feedback and continually strive to improve our products and services. 💬

Sustainability:

- Committed to sustainability, Pure Fresh implements eco-friendly practices throughout its operations.
- We support local farmers, promote animal welfare, and minimize our environmental footprint. 🌍

Looking Ahead:

As Pure Fresh looks to expand, we are seeking partnerships and opportunities to bring our fresh, high-quality dairy products to a wider audience. We are excited about the potential for growth while staying true to our mission and values. 🚀

Contact Information:

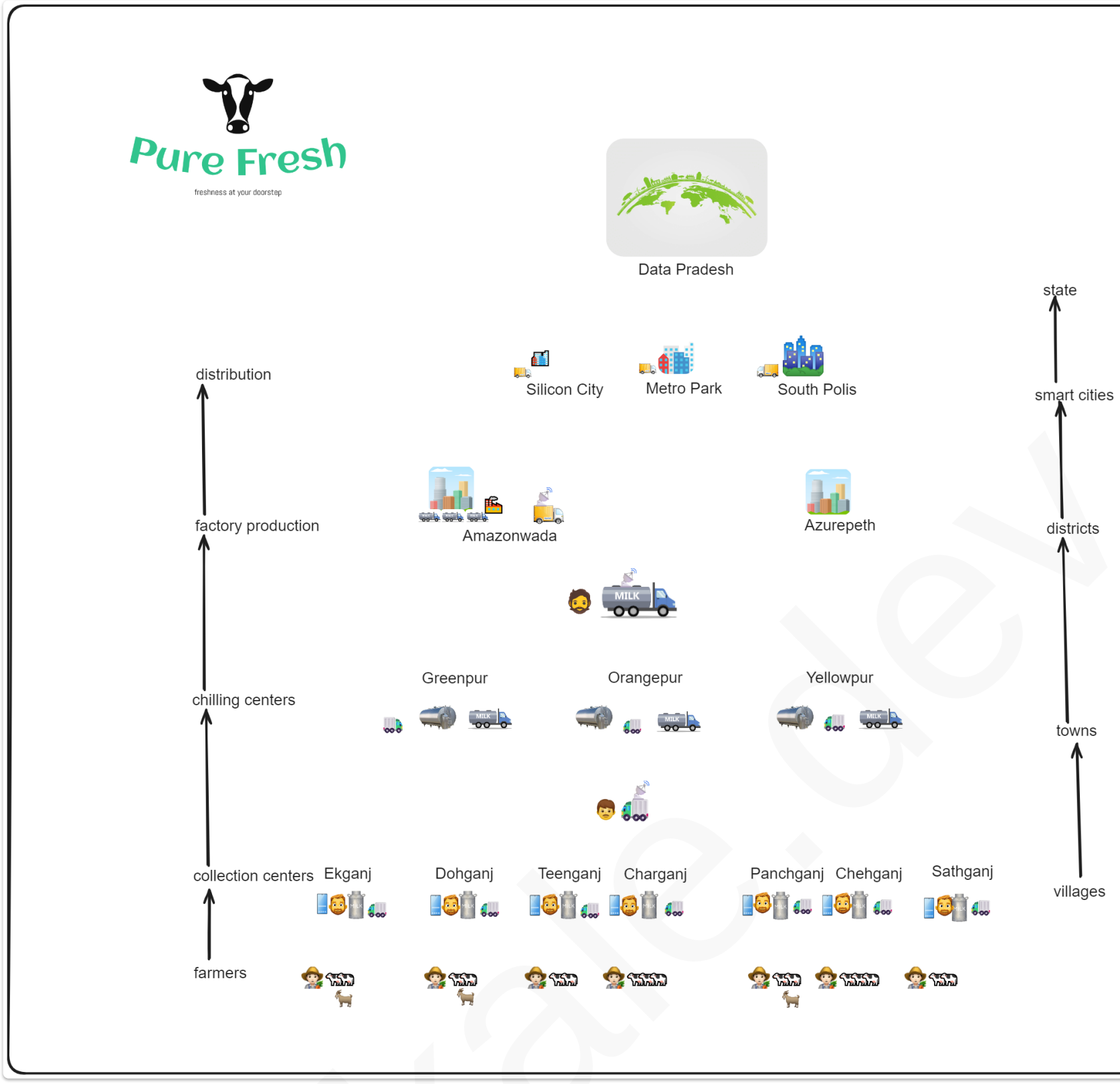
- Website: www.purefreshdairy.com
- Email: info@purefreshdairy.com
- Phone: 1-800-PUREMILK
- (Obviously, the above info is made up 😉)

How Does Pure Fresh Operate?

Pure Fresh sources milk from farmers and creates products, which they distribute in the market. Let's understand how the business flows in brief.

Firstly, the farmers bring the milk to collection centers. At the centers, the center owners test the quality of the milk. The milk data is fed into an app by the center owners. Independent contractors then collect milk from centers and take it to chilling centers. At the chilling center, milk from various collection centers is aggregated, and the quality of each center is evaluated. After that, the aggregated milk is taken to the factory, where QA is done at the chilling center level again. In the factory, five types of products are created. These products are then delivered to the distributors, who sell them further. 🏭

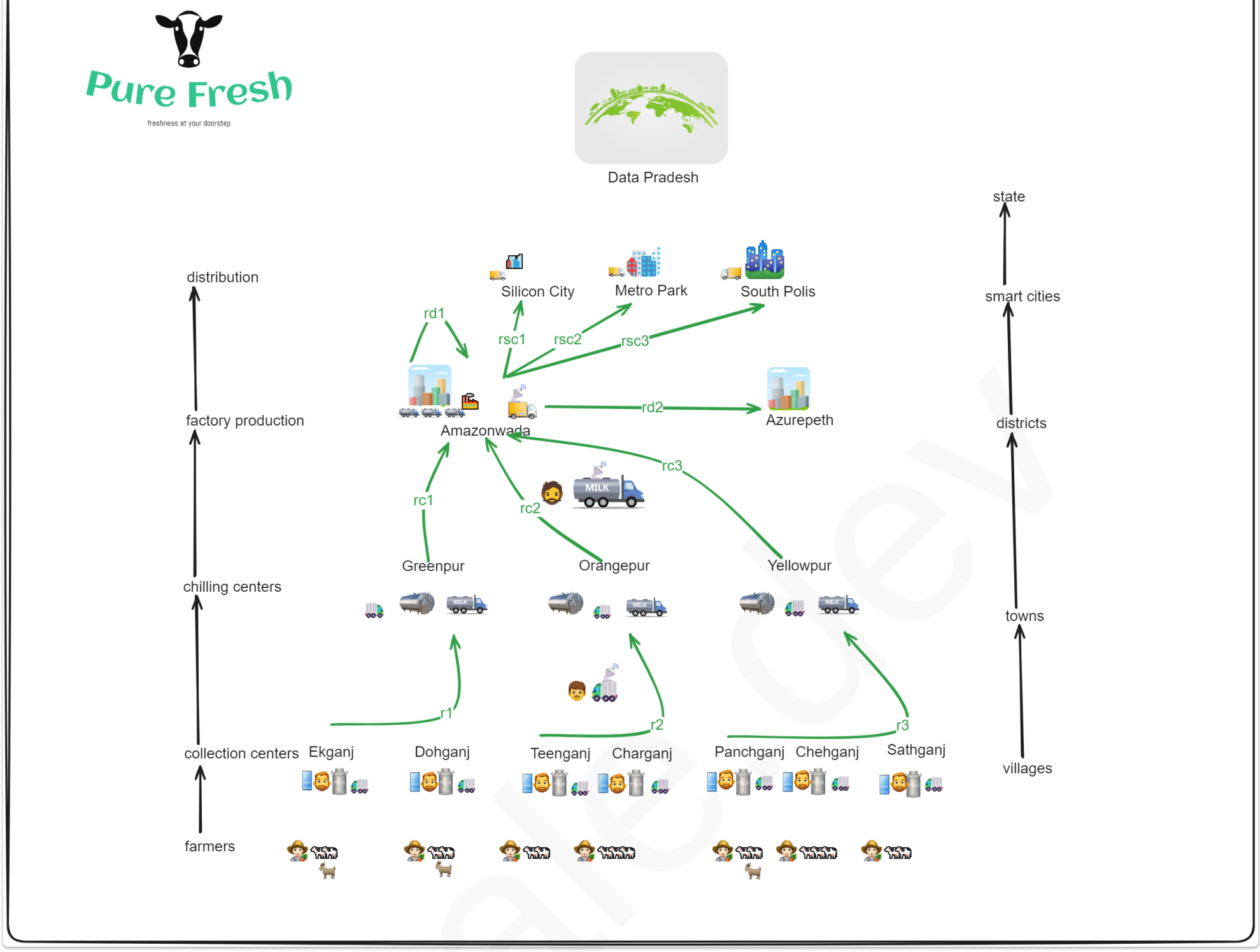
So that's the business overall. Have a look at the visualization below for more clarity.



The routes for milk collection are as shown below.

Note

The number of farmers, collection centers, chilling centers, and routes are static.
In a nutshell, the dim prospects are kept static purposefully while coding for data generation.
Only the potential facts are generated for a given date.



How is Data Generated at the Business Level?

So, Pure Fresh has data of two types: static or slowly changing data and rapidly changing data.

> **Remember**
>
> As I have generated the data, the data generation will only produce the rapidly changing data for the given date. The slowly changing data is kept as is.
>
> There are some limitations for mock data generation, and keeping the business context, some things are hard-set, like the number of collection centers, the chilling centers, etc.

Slowly Changing Data

This is the data that is collected and stored in their system. This data changes infrequently.

- farmers_data:
  - Data about farmers, like their details.
- collection_center_data:
  - Details of collection centers, the owners, and other specifics.
- chilling_center_data:
  - Details about chilling centers, the owners, and other specifics.
- logistics_contractor_data:
  - Data about the contractors, their vehicle numbers, etc.
- routes_data:
  - Details about the route, the stops, and the distance, etc.

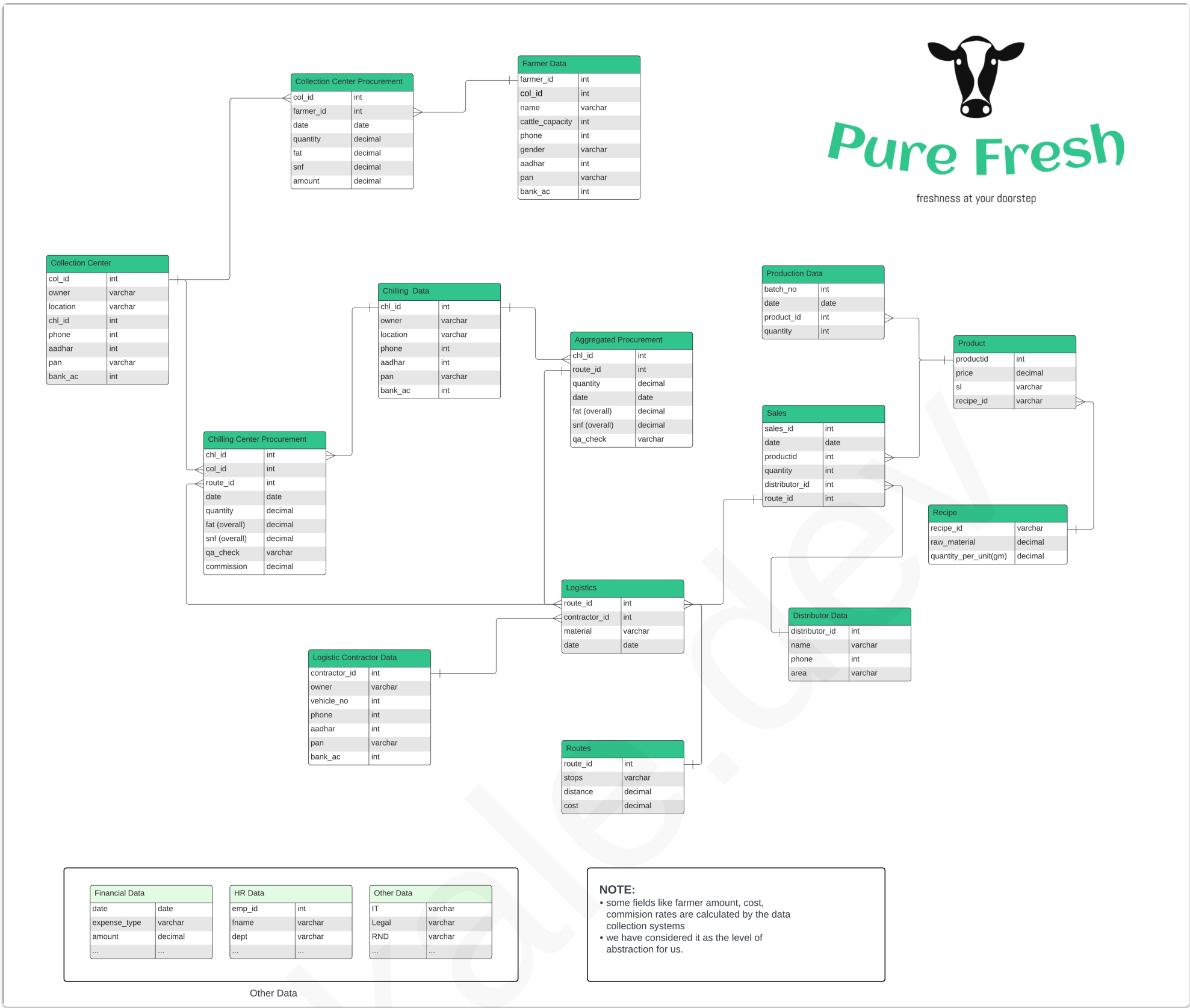
- distributor_data:
  - Details about the distributors.
- products_data:
  - Details about the products, etc.
- recipe_data:
  - Information about the raw materials required for the products.

Rapidly Changing Data

This data is generated on a daily basis.

- collection_center_procurement:
  - Data about daily milk sourced from a farmer, collected at the collection center level.
- chilling_center_procurement:
  - Data about the total milk sourced from a collection center daily, collected at the chilling center level.
- aggregated_procurement:
  - Data about the total milk sourced from a chilling center daily, at the factory level.
- daily_logistics:
  - Data about the logistics contractors and routes, detailing which contractor picked milk from each location and other specifics.
- production_data:
  - Data about the quantity of total products manufactured on a daily basis.
- sales_data:
  - Data about how many products were sold to which distributor.

Below is the data model of Pure Fresh:



What Are the Pain Points of Pure Fresh?

Pure Fresh is doing business in a small territory and is planning for expansion. To do that, they want to understand their current state and use data points to support their decisions. For example, how to introduce new products without affecting current products? What are the current pain points at all levels? 🤔

The BA team at Pure Fresh has identified certain points that need to be controlled:

- QA at all levels
- Farmers need credit to produce more milk
- Pure Fresh needs to attract more farmers to source milk for them, ensuring methods to attract new farmers.
- Need to minimize loss in logistics. 🚚

What's the Solution?

Pure Fresh is looking for help to understand the patterns and rationally make business decisions. As a data solution, they are looking for something that will aggregate all their data in one place. They then want some reports to be generated from this data. The solution should consider modern data architecture and technologies. As Pure Fresh is still in its initial phase, they are looking for a solution that is light on budget, both in terms of infrastructure and resources. The solution can be divided into phases, with the first one being creating reports for existing data. 📊

That's it. I hope you understood the problem statement. The sample data is here. I will see you with some progress in the next step. 🚀