Requirements:

1. You are hired by cooking.com and they need to show a page with list of all the recipes and when user clicks on each of the recipe they want to show the Recipe page with their ingredients. They also want user to further click into each of the ingredient and see all the recipes linked to that ingredients.

You are hired as a Data Developer to make data available for Frontend systems to consume which will make this possible. Currently, recipes data is received as feeds from legacy system in a form of a CSV. CSV data looks like this

**recipe\_id, recipe\_name, description, ingredient, active, updated\_date, created\_date**

1, pasta, Italian pasta, tomato sauce, true, 2018-01-09 10:00:57, 2018-01-10 13:00:57

1, pasta, null, cheese, true, 2018-01-09 10:10:57, 2018-01-10 13:00:57

2, lasagna, layered lasagna, cheese, true, 2018-01-09 10:00:57, 2018-01-10 13:00:57

2, lasagna, layered lasagna, blue cheese, false, 2018-01-09 10:00:57, 2018-01-10 13:00:57 ….

Assume that this CSV is consumed every 1 hour with 1TB of data You are asked to:

1. Create a data model which can store this data to allow user to do the activities mentioned above. This data model needs to support millions of read per second.
2. Discuss the persistence system you are going to use to store this data.
3. Write a Spark Job in Scala which can takes the CSV shown above and store that in the storage system of your choice using the data model you discussed above.
4. Write queries to answer the following
   1. Average number of recipes which are updated per hour
      1. Eg. Pasta got updated twice in one hour
   2. Number of recipes which got updated at 10:00 clock in the entire year.

Requirements summary:

Design system that will support web application that displays Cooking recipes to users and provides regular data updates and performs data analytic jobs.

In requirement:

“Write queries to answer the following

* Average number of recipes which are updated per hour

Eg. Pasta got updated twice in one hour”

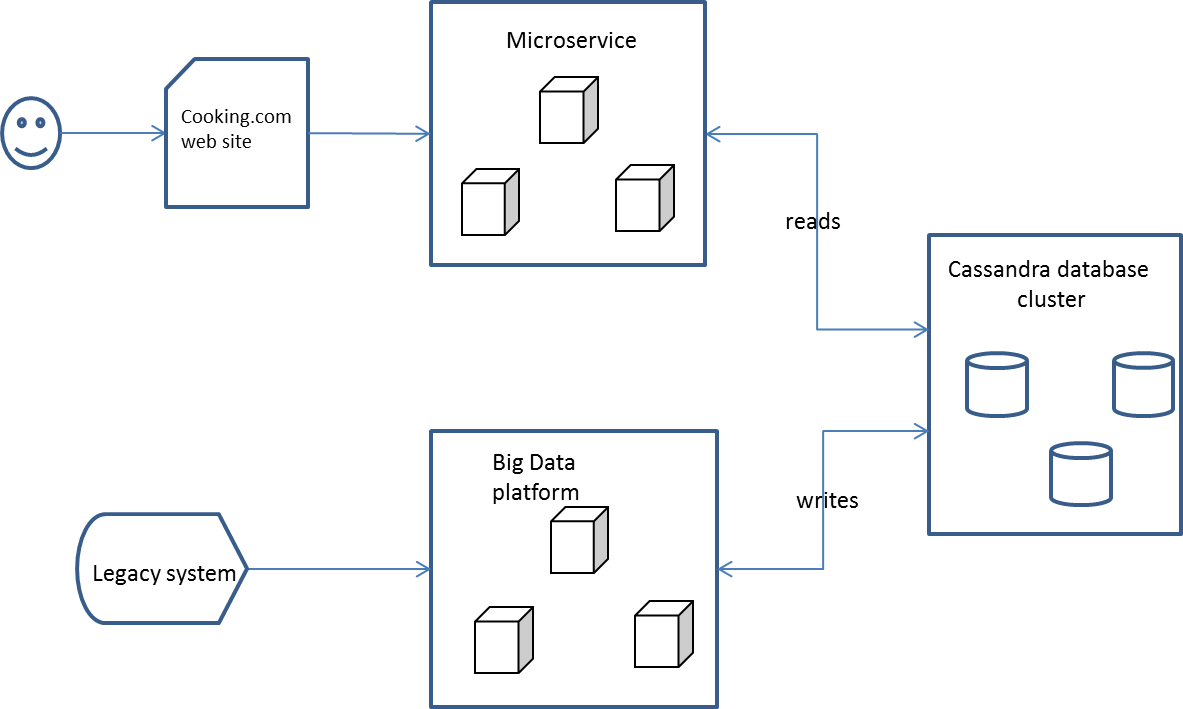
Actual requirement doesn’t correspond to the provided example. Requirement asks for average number of updated recipes but not for how many times each individual recipe was updated each hour, according to example. Implementation is done based on requirement.

Assumptions:

1. File somehow is uploaded to Hadoop cluster and saved in chunks on multiple nodes
2. List of ingredients for recipe is not changeable only active state is changed. Adding new ingredient to a recipe would mean a new recipe

System design:

The following diagram represents high level architecture of the proposed solution.



It consists from three main components:

* Web application backend – web application backend will be built using Reactive Microservices architecture using Akka. Scala Lagom microservices framework is a good candidate to implement this component. It supports asynchronous communication and has built in support for Cassandra database.
* Big Data platform cluster – any Hadoop base Big Data platform (Cloudera, Hortonworks)
* Cassandra database cluster – database component will be built with Cassandra database. Cassandra database selected for the following reasons:
* NoSQL column oriented database
* Provides SQL like query support
* Cassandra allows for a primary key to contain multiple columns
* Cassandra has a masterless architecture and provides high availability
* Provides very fast write operations, it’s an efficient write-oriented database.

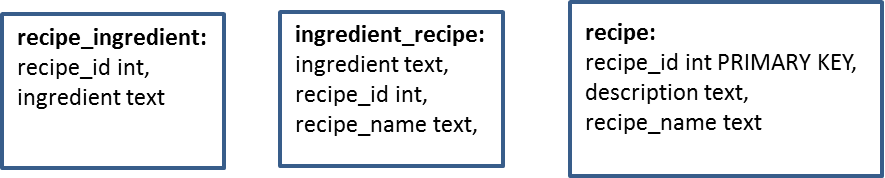
Data model design:

Data model is design around Event Sourcing and CQRS principles. Data will be stored in three tables:

The main feed from the CSV file will be saved in recipe\_event table and this table will be used for all data analytic queries.



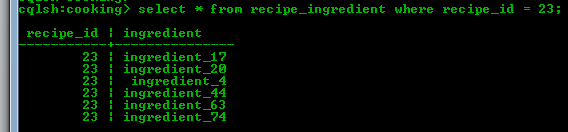
Tables: recipe, ingredient\_recipe and recipe\_ingredient will be used to server requests from web application.



1. Get list of recipes – table recipe will be used



1. Get ingredient for selected recipe - table recipe\_ingredient will be used



1. Get list of recipes for selected ingredient – table ingredient\_recipe will be used

