Cassandra – Data Modeling

1 Introduction

Data modeling in Cassandra is one of the most important factors in designing your cluster. A properly designed data model allows Cassandra to support massive, linear scalability.

1.1 Physical vs. logical model

When modeling, focus on the physical storage as it will determine the performance characteristics of your queries. Do not think about the logical CQL abstraction when modeling.

> **Tip**: Physical, on-disk storage will determine the quality of your data model.

1.1.1 CQL

Let’s begin by creating a simple CQL table that stores the population of various cities.

> **Hint**: The clustering columns specified via `PRIMARY KEY` allow a partition to have multiple rows.

```
CREATE TABLE population (
    country_id uuid,
    state text,
    city text,
    population int,
    PRIMARY KEY (country_id, state, city)
);
```

1.1.2 CQL logical representation

CQL logically displays a single partition as multiple rows in a traditional tabular format.

> **Warning**: Design your data model based on physical storage, not CQL's abstraction.

One of the first things you’ll notice below is that the partition key component of the primary key repeats for each CQL row.

country_id	state	city	population
US	CA	SD	2598756
US	CA	SF	3258489