# Introduction

### Background Information

The project aims to investigate the property prices of Beijing from 2011 to 2017 and gain insights into the market trends. Data is stored using HDFS (Hadoop Distributed File System) and analysed using Apache Spark in Jupyter Notebook with Python.

### Motivation

Property market plays a vital role in China’s economy, Beijing, being both the capital and a major economic hub, its property price often reflect the broader economic trend in the rest of China.

The city has been the testing ground for several real estate policies aimed at controlling property prices, analysing prices in Beijing can help shed light on the effectiveness of these policies.

Beijing’s housing market also attracts both domestic and international investors, understanding the price trend over time can provide investors with valuable insights into the market.

### Objective/Features

The project seeks to investigate 4 issues:

* Average house price over the years.
  + Provides a view of the overall market, crucial for identifying economic growth and slowdowns.
* Change in the price per square by district over the years.
  + Help identify districts which have experienced the most growth and the best outlook, important for investors when buying a property.
* Change in house specification over the years.
  + Reflect shift in consumer preference, whether there are more people living alone, change in family size and more.
* Change in price by districts over the years.
  + Whether consumers are concentrating in certain regions, such that decline in the overall market is not reflected here. This is important for both buyers and investors looking to compare growth rates and make informed decisions.

### Limitations of traditional computing

Traditional computing relies on a single machine, which has limitations in scaling both processing power and storage. The reliance on one machine makes the applications more prone to failure, having no ways of backup.

### Benefits of cloud computing

Cloud computing fits well with the project as it provides a way of scaling in the future easily and cost-effectively.

Data can be processed in parallel across multiple machines, speeding up the processing time for large datasets. Increase in data over the years can be solved easily by adding more machines.

The machines are also cheaper and have lower specification requirements instead of expensive high-end traditional storage systems.

Multiple copies of the datasets are available spread across the machines, ensuring the project is fault tolerant.

# Technical Solutions

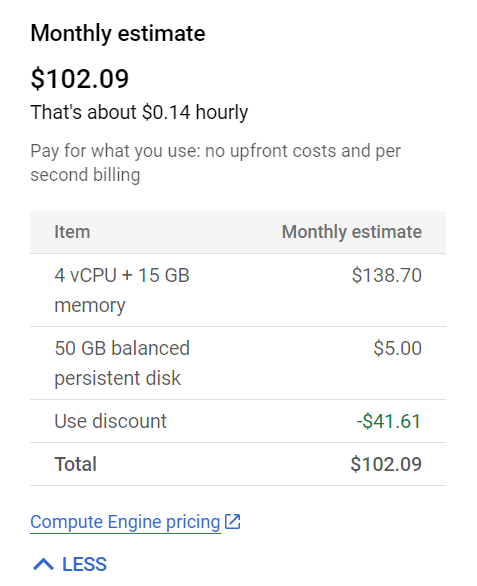
### Cloud technologies used

VM (virtual machines) used to easily scale the computing resources based on needs, such as adding more CPUs, allocating more memory, changing operating systems. VM also allows application to be run in isolated environment, so any crash does not affect the host machine and can revert to previous snapshots if needed.

HDFS used ensure the data availability and scalability. It is designed to handle large datasets by distributing the data across multiple nodes, keeping multiple copies, preventing data loss and maintaining system availability. Three HDFS nodes are used for the project to ensure data availability.

Apache Spark used for speed and scalability, able to scale from a single machine into thousands of nodes, capable of handling large datasets. The project uses two Spark worker nodes to spread the computing power, increasing the processing speed of the project.

### Monthly cost estimates



# Architecture Design

Data exploration, Data cleaning, Data analysis, Results discussion