بسم اللّٰه الرحمن الرحيم

Apache Drill

امیرحسین بنایی خلیل آباد

درس مباحث ویژه ۱

نام استاد

ابوالفضل گندمی

What is Apache Drill

- Drill is an Apache open-source SQL query engine for Big Data exploration. Drill is designed from the ground up to support high-performance analysis on the semi-structured and rapidly evolving data coming from modern Big Data applications, while still providing the familiarity and ecosystem of ANSI SQL, the industry-standard query language. Drill provides plug-and-play integration with existing Apache Hive and Apache HBase deployments.

Apache Drill Key Features

- Low-latency SQL queries
- Dynamic queries on self-describing data in files (such as JSON, Parquet, text) and HBase tables, without requiring metadata definitions in the Hive metastore.
- ANSI SQL
- Nested data support
- Integration with Apache Hive (queries on Hive tables and views, support for all Hive file formats and Hive UDFs)

What is NoSQL?

- NoSQL databases (aka "not only SQL") are non-tabular databases and store data differently than relational tables. NoSQL databases come in a variety of types based on their data model. The main types are document, key-value, wide-column, and graph. They provide flexible schemas and scale easily with large amounts of data and high user loads.

Student Table in SQL Server

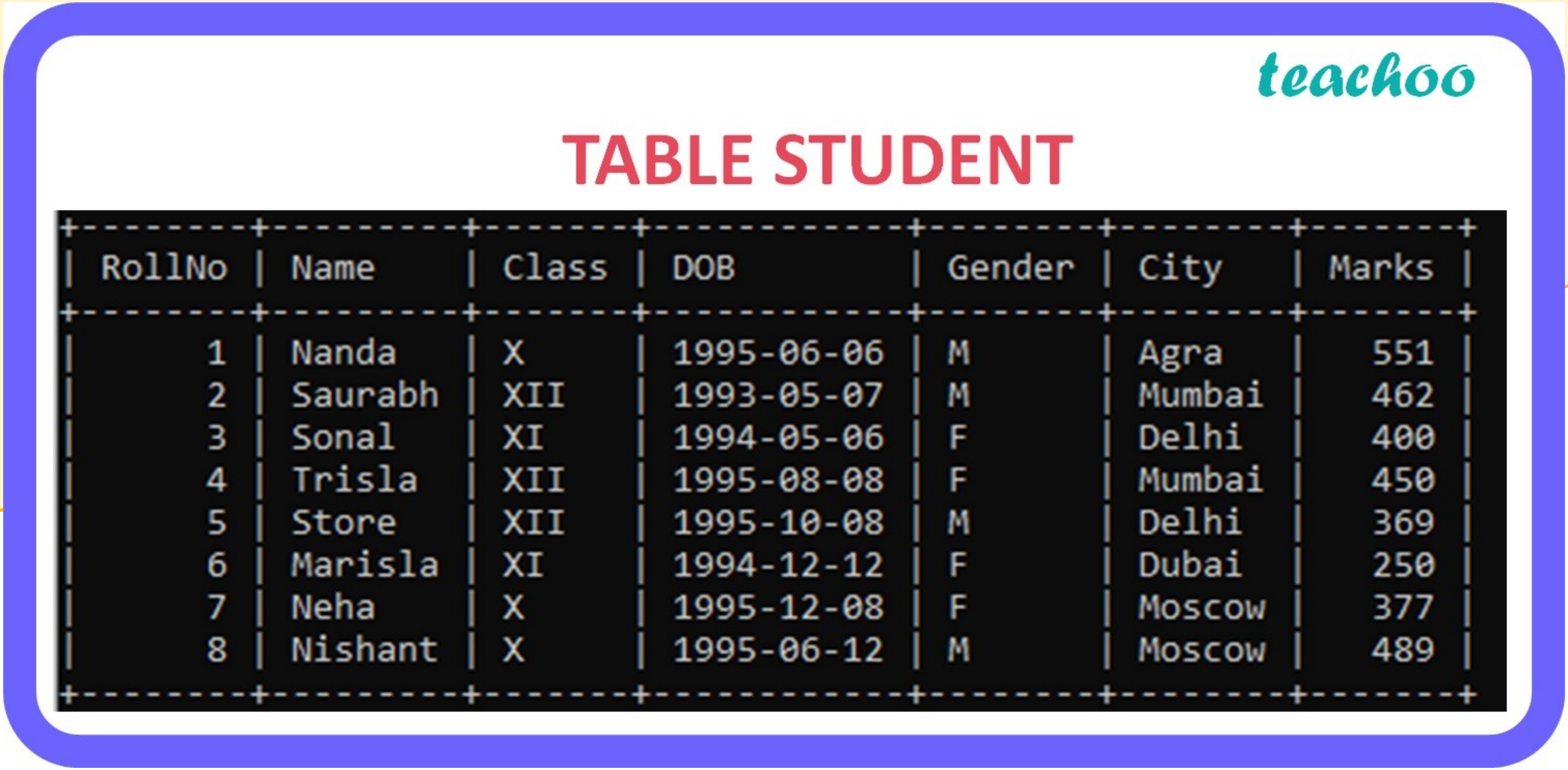
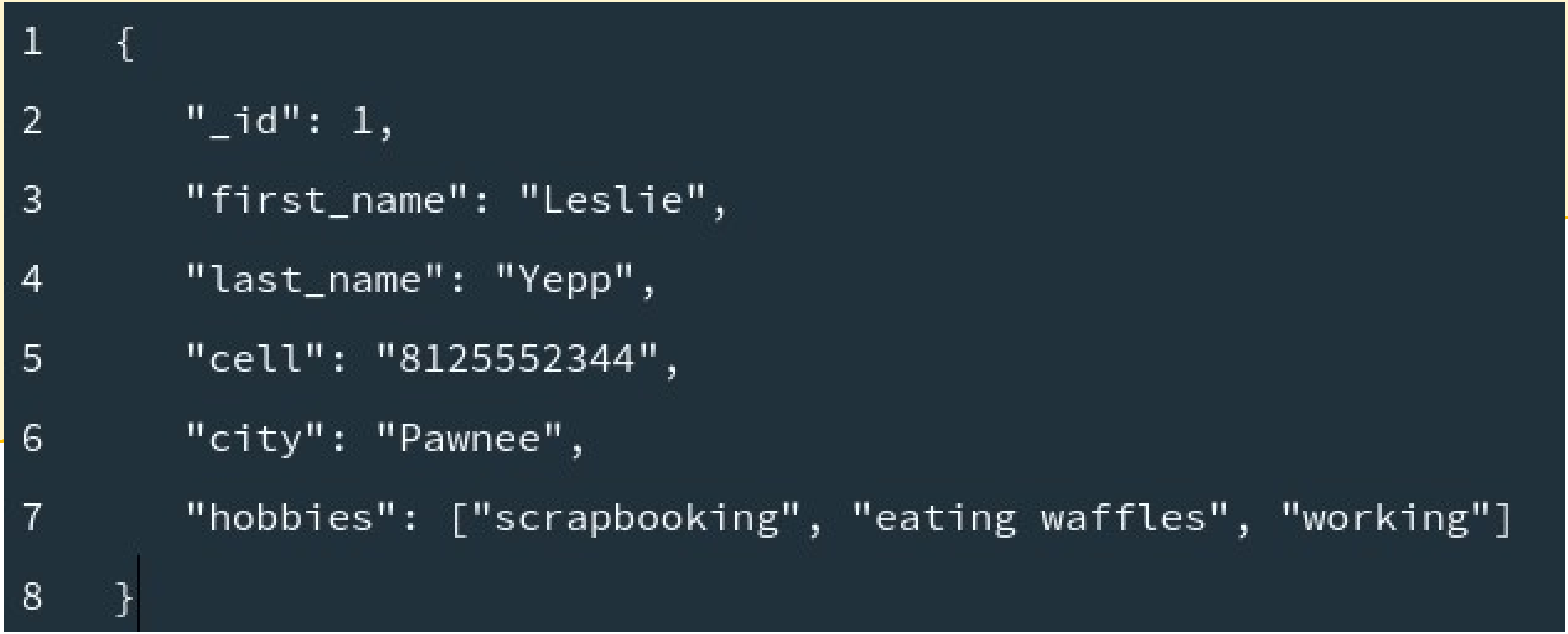


TABLE STUDENT

```
+--------+----------+-------+------------+---------+--------+-------+
| RollNo | Name     | Class | DOB        | Gender  | City   | Marks |
+--------+----------+-------+------------+---------+--------+-------+
|      1 | Nanda    | X     | 1995-06-06 | M       | Agra   |   551 |
|      2 | Saurabh  | XII   | 1993-05-07 | M       | Mumbai |   462 |
|      3 | Sonal    | XI    | 1994-05-06 | F       | Delhi  |   400 |
|      4 | Trisla   | XII   | 1995-08-08 | F       | Mumbai |   450 |
|      5 | Store    | XII   | 1995-10-08 | M       | Delhi  |   369 |
|      6 | Marisla  | XI    | 1994-12-12 | F       | Dubai  |   250 |
|      7 | Neha     | X     | 1995-12-08 | F       | Moscow |   377 |
|      8 | Nishant  | X     | 1995-06-12 | M       | Moscow |   489 |
+--------+----------+-------+------------+---------+--------+-------+
```

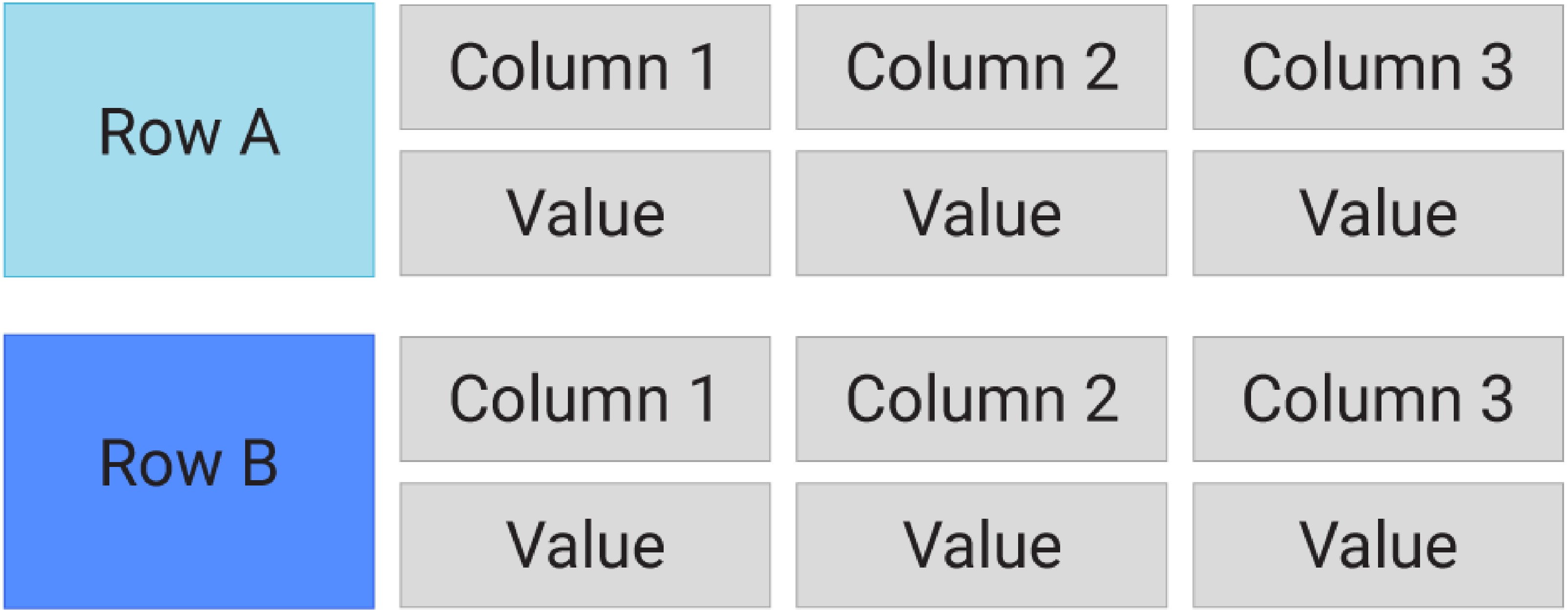
Student Table in SQL Server

JSON format in NoSQL



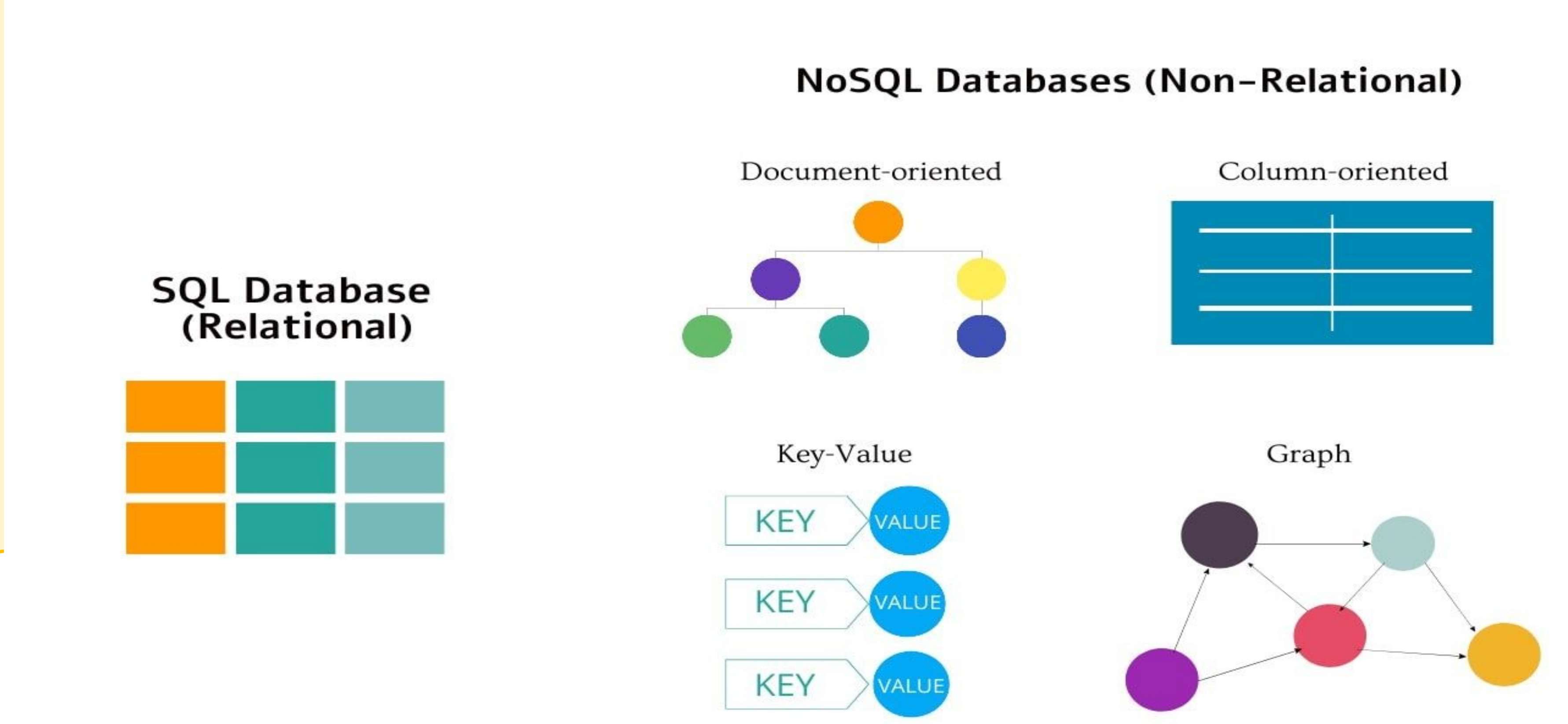
```
{
    "_id": 1,
    "first_name": "Leslie",
    "last_name": "Yepp",
    "cell": "8125552344",
    "city": "Pawnee",
    "hobbies": ["scrapbooking", "eating waffles", "working"]
}
```

What is wide-column in NoSQL?

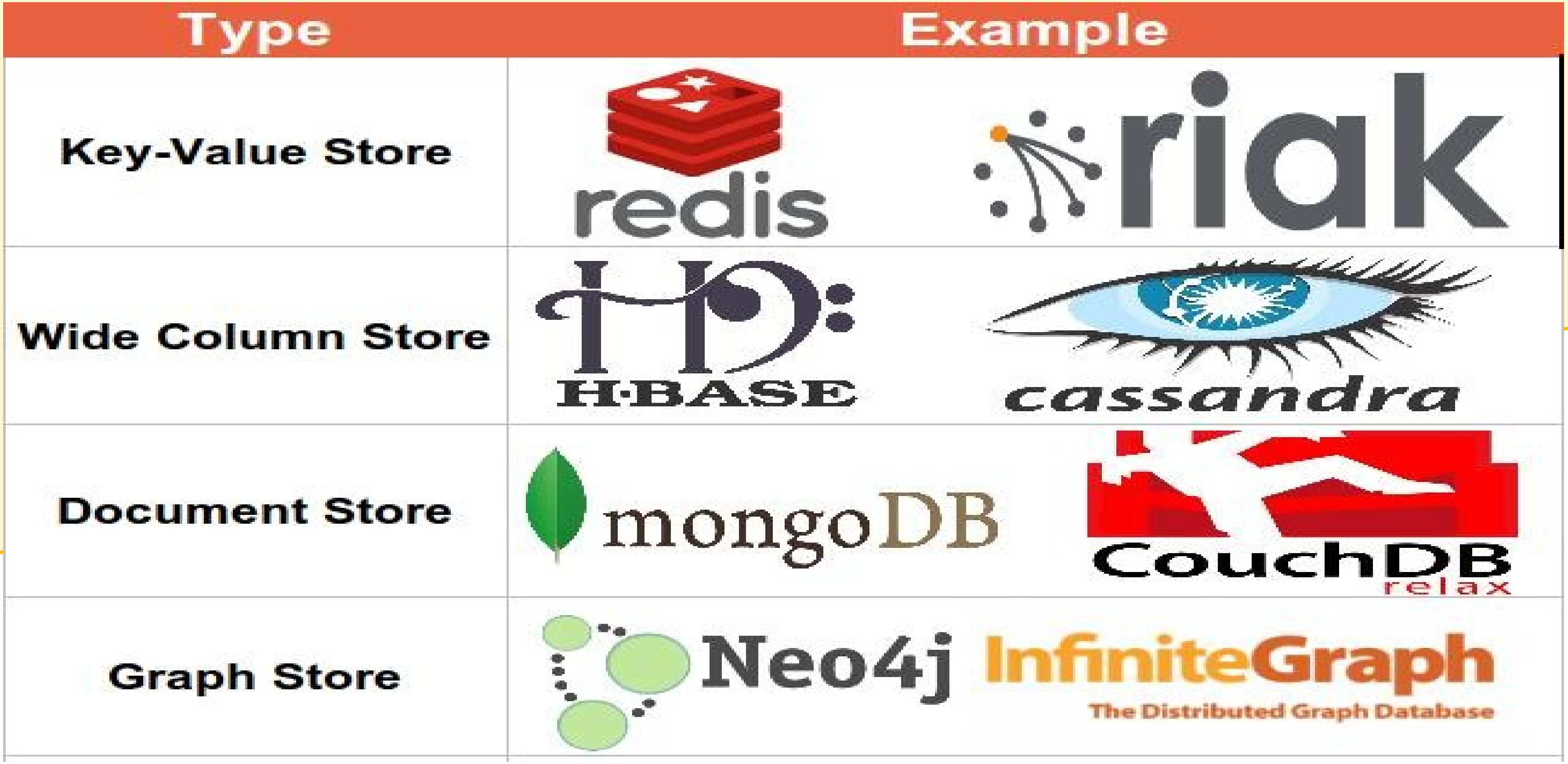


Row A	Column 1	Column 2	Column 3
	Value	Value	Value
Row B	Column 1	Column 2	Column 3
	Value	Value	Value

What is graph in NoSQL?

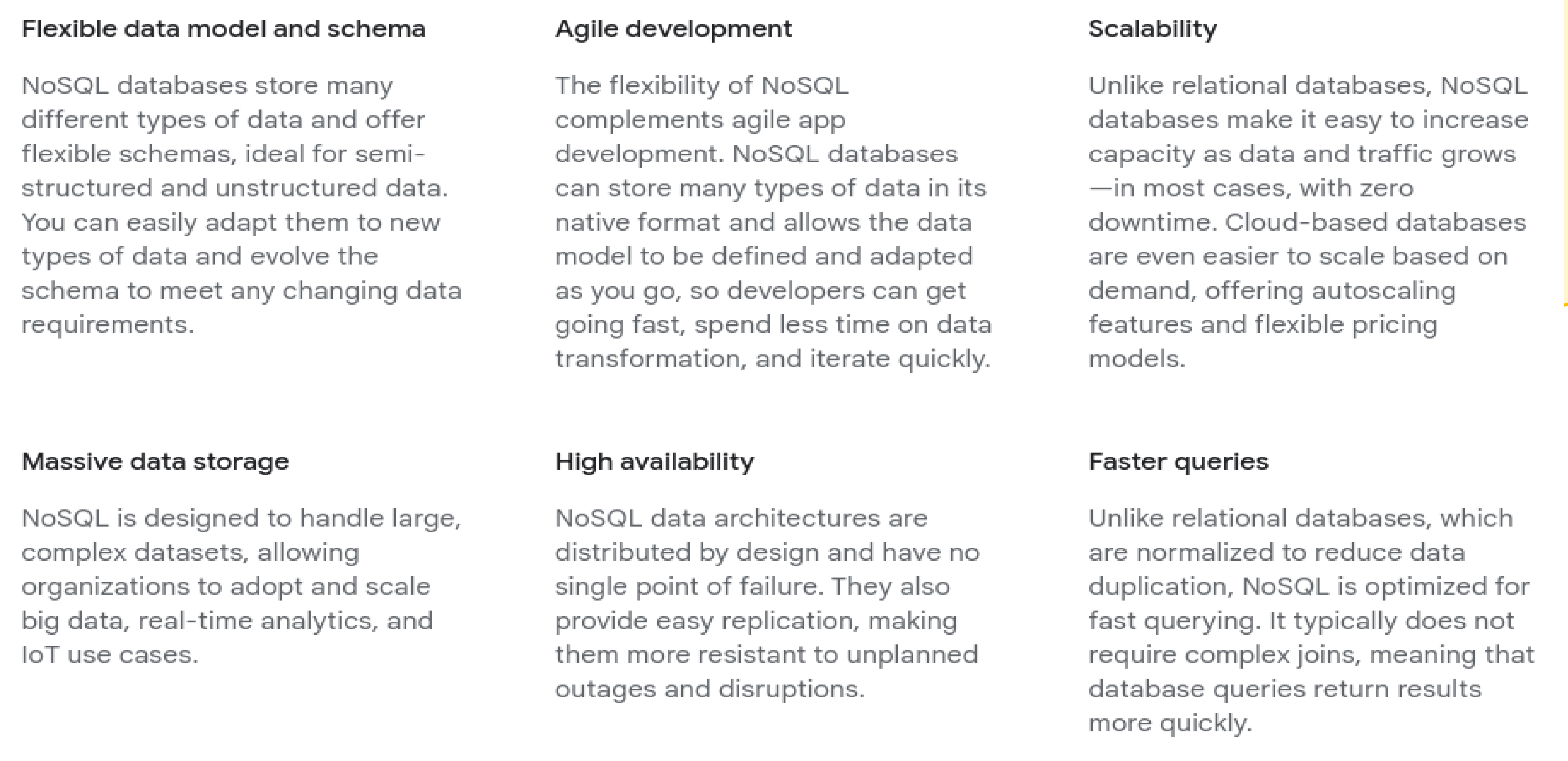


NoSQL



Type	Example	
Key-Value Store	redis	riak
Wide Column Store	H·BASE	cassandra
Document Store	mongoDB	CouchDB relax
Graph Store	Neo4j	InfiniteGraph The Distributed Graph Database

Why NoSQL?



Flexible data model and schema

NoSQL databases store many different types of data and offer flexible schemas, ideal for semi-structured and unstructured data. You can easily adapt them to new types of data and evolve the schema to meet any changing data requirements.

Agile development

The flexibility of NoSQL complements agile app development. NoSQL databases can store many types of data in its native format and allows the data model to be defined and adapted as you go, so developers can get going fast, spend less time on data transformation, and iterate quickly.

Scalability

Unlike relational databases, NoSQL databases make it easy to increase capacity as data and traffic grows —in most cases, with zero downtime. Cloud-based databases are even easier to scale based on demand, offering autoscaling features and flexible pricing models.

Massive data storage

NoSQL is designed to handle large, complex datasets, allowing organizations to adopt and scale big data, real-time analytics, and IoT use cases.

High availability

NoSQL data architectures are distributed by design and have no single point of failure. They also provide easy replication, making them more resistant to unplanned outages and disruptions.

Faster queries

Unlike relational databases, which are normalized to reduce data duplication, NoSQL is optimized for fast querying. It typically does not require complex joins, meaning that database queries return results more quickly.

Install Drill Introduction

1: embedded mode

2 ways

2: distributed mode

Install Drill Introduction on embedded mode

- You can install Drill for use in either embedded mode or distributed mode. Choose embedded mode to use Drill only on a single node. Installing Drill for use in embedded mode does not require installation of ZooKeeper. Using Drill in embedded mode requires no configuration.

Install Drill Introduction on distributed mode

- Choose distributed mode to use Drill in a clustered Hadoop environment. A clustered (multi-server) installation of ZooKeeper is one of the prerequisites. You also need to configure Drill for use in distributed mode. After you complete these tasks, connect Drill to your Hive, HBase, or distributed file system data sources, and run queries on them.

Embedded Mode Prerequisites

Before you install Drill, ensure that the machine meets the following prerequisites:

- Linux, Mac OS X, and Windows: Oracle or OpenJDK 8
- Windows only:
- A JAVA_HOME environment variable that points to the JDK installation
- A PATH environment variable that includes a pointer to the JDK installation
- A utility for unzipping a tar.gz file

Running Drill on Docker

- You can start and run a Docker container in detached mode or foreground mode. Detached mode runs the container in the background. Foreground is the default mode. Foreground mode runs the Drill process in the container and attaches the console to Drill’s standard input, output, and standard error.

Running the Drill Docker Container in Foreground Mode

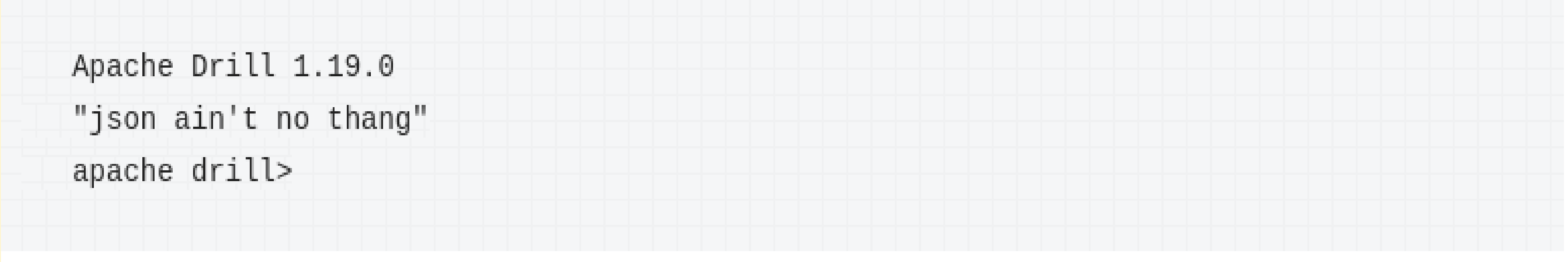
-
  ```
  docker run -it --name drill \
  -p 8047:8047 \      # web and REST
  -p 31010:31010 \    # JDBC
  apache/drill
  ```

Running the Drill Docker Container in Detached Mode

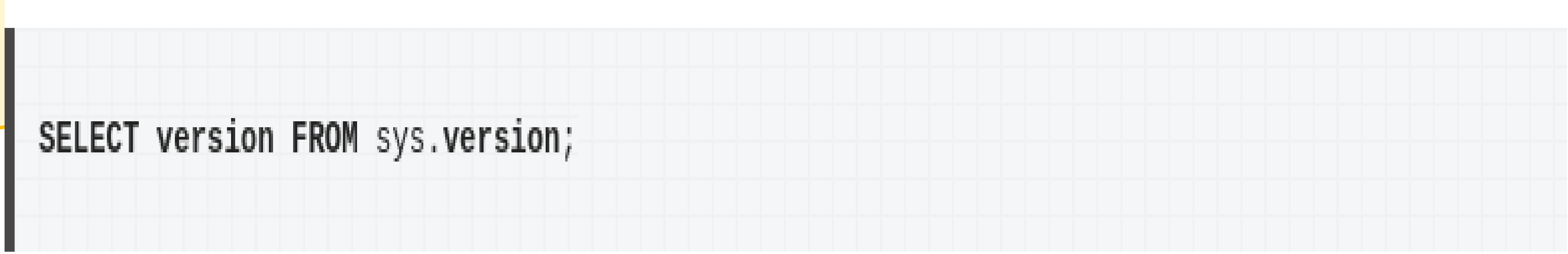
-
  ```
  $ docker run --name drill \
  -p 8047:8047 \       # web and REST
  -p 31010:31010 \   # JDBC
  --detach
  apache/drill

  $ docker exec -it drill /bin/bash
  $ $DRILL_HOME/bin/drill-embedded
  ```

Confirm Install Drill

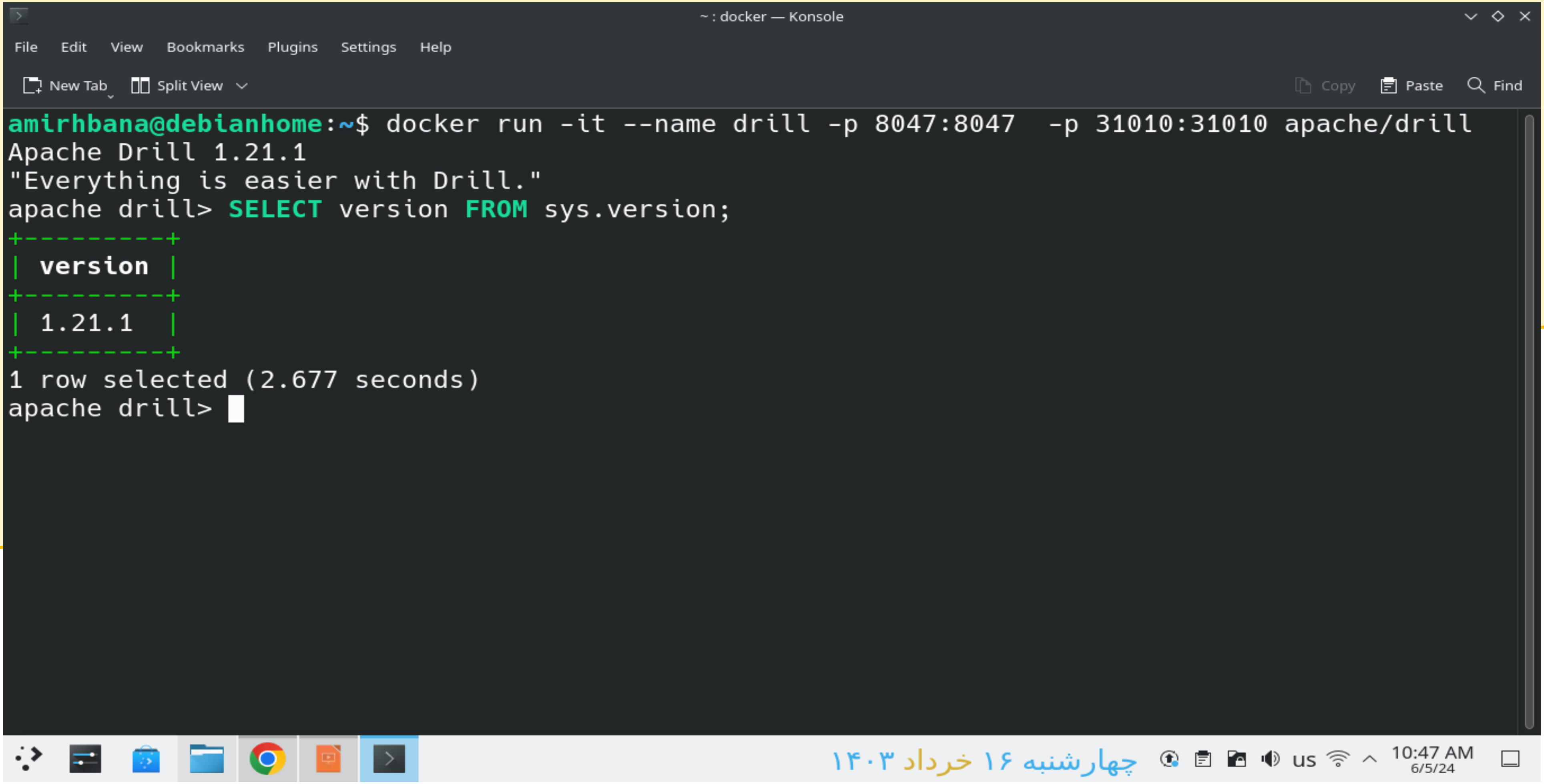


```
Apache Drill 1.19.0
"json ain't no thang"
apache drill>
```



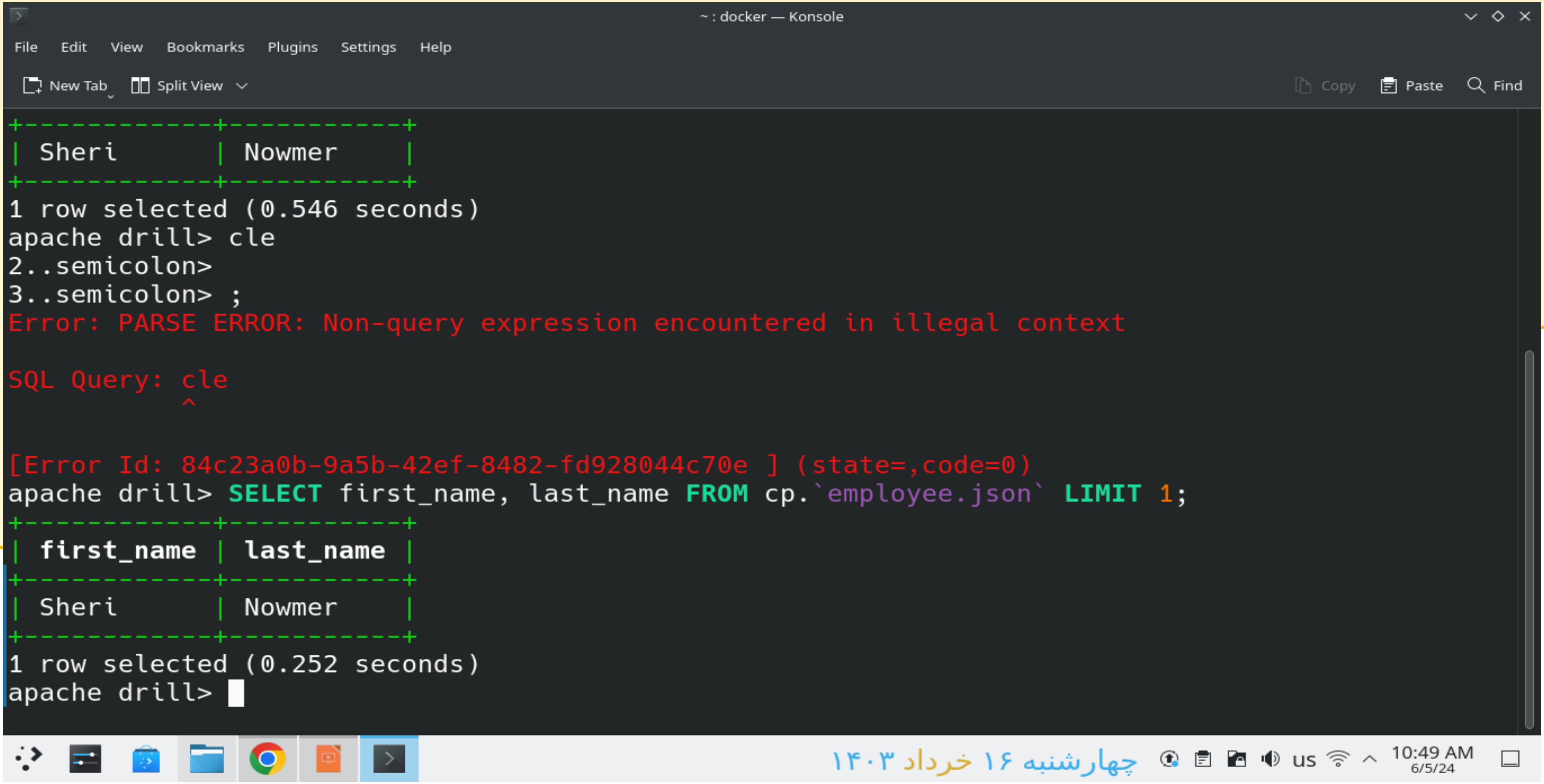
```
SELECT version FROM sys.version;
```

Confirm Install Drill



```
amirhbana@debianhome:~$ docker run -it --name drill -p 8047:8047  -p 31010:31010 apache/drill
Apache Drill 1.21.1
"Everything is easier with Drill."
apache drill> SELECT version FROM sys.version;
+----------+
| version |
+----------+
| 1.21.1  |
+----------+
1 row selected (2.677 seconds)
apache drill> 
```

SELECT first_name, last_name FROM cp.`employee.json` LIMIT 1;



```
+------------+------------+
| Sheri      | Nowmer     |
+------------+------------+
1 row selected (0.546 seconds)
apache drill> cle
2..semicolon>
3..semicolon> ;
Error: PARSE ERROR: Non-query expression encountered in illegal context

SQL Query: cle
           ^


[Error Id: 84c23a0b-9a5b-42ef-8482-fd928044c70e ] (state=,code=0)
apache drill> SELECT first_name, last_name FROM cp.`employee.json` LIMIT 1;
+------------+------------+
| first_name | last_name  |
+------------+------------+
| Sheri      | Nowmer     |
+------------+------------+
1 row selected (0.252 seconds)
apache drill>
```