

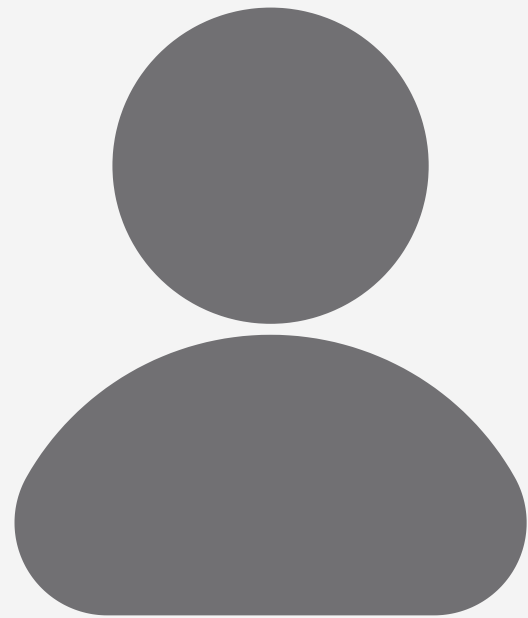
BÁO CÁO SEMINAR CHƯƠNG 3



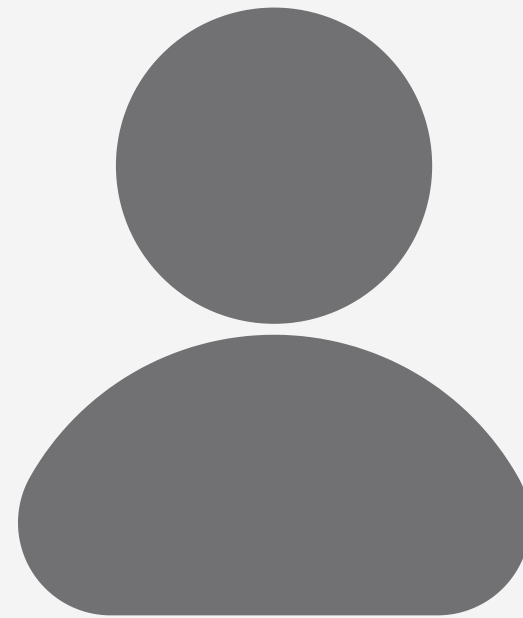
APACHE DRILL

DỮ LIỆU LỚN – IS405.P11.HTCL
GVHD: ThS. Nguyễn Hồ Duy Tri

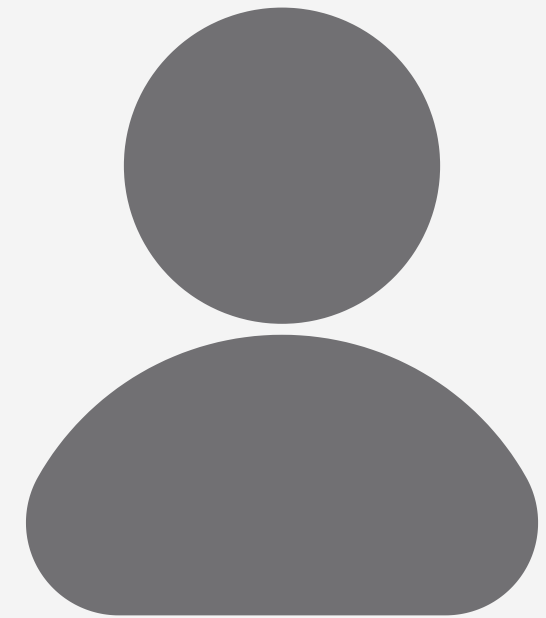
THÀNH VIÊN



Trương Vĩnh Thuận
21522653



Hoàng Quốc Việt
21522790



Phùng Thiên Phúc
21521297

THÔNG TIN CHUNG

1. Khái niệm

Drill là một công cụ truy vấn SQL mã nguồn mở của Apache, được thiết kế để phân tích dữ liệu lớn một cách hiệu quả. Nó hỗ trợ phân tích dữ liệu bán cấu trúc và phát triển nhanh từ các ứng dụng Big Data hiện đại, đồng thời sử dụng ngôn ngữ truy vấn ANSI SQL quen thuộc. Drill có thể tích hợp dễ dàng với Apache Hive và Apache HBase.

THÔNG TIN CHUNG

2. Tính năng nổi bật

- Truy vấn SQL độ trễ thấp
- Truy vấn trên nhiều loại dữ liệu
- Hỗ trợ ANSI SQL
- Hỗ trợ dữ liệu lồng ghép (nested data)
- Tích hợp với Apache Hive
- Tích hợp công cụ BI/SQL

ƯU ĐIỂM VÀ NHƯỢC ĐIỂM

ƯU ĐIỂM

1. Hỗ trợ truy vấn dữ liệu phi cấu trúc và bán cấu trúc
2. Tính linh hoạt trong việc kết nối với nhiều hệ thống lưu trữ
3. Dễ sử dụng với SQL quen thuộc
4. Khả năng mở rộng và phân tán
5. Không cần ETL và khả năng thời gian thực
6. Cộng đồng và phát triển nguồn mở

NHƯỢC ĐIỂM

1. Hiệu năng có thể không ổn định với các tập dữ liệu rất lớn
2. Chưa hỗ trợ tốt các tác vụ phân tích chuyên sâu
3. Yêu cầu cấu hình và quản lý cụm (cluster) phức tạp
4. Giới hạn trong hỗ trợ ACID và bảo mật

ĐẶC ĐIỂM NỔI BẬT

1. Hỗ trợ schema-free
2. Truy vấn nhiều nguồn dữ liệu từ một hệ thống
3. Cộng đồng phát triển mạnh mẽ
4. Khả năng xử lý dữ liệu bán cấu trúc tốt hơn

SO SÁNH VỚI SẢN PHẨM CÙNG LOẠI



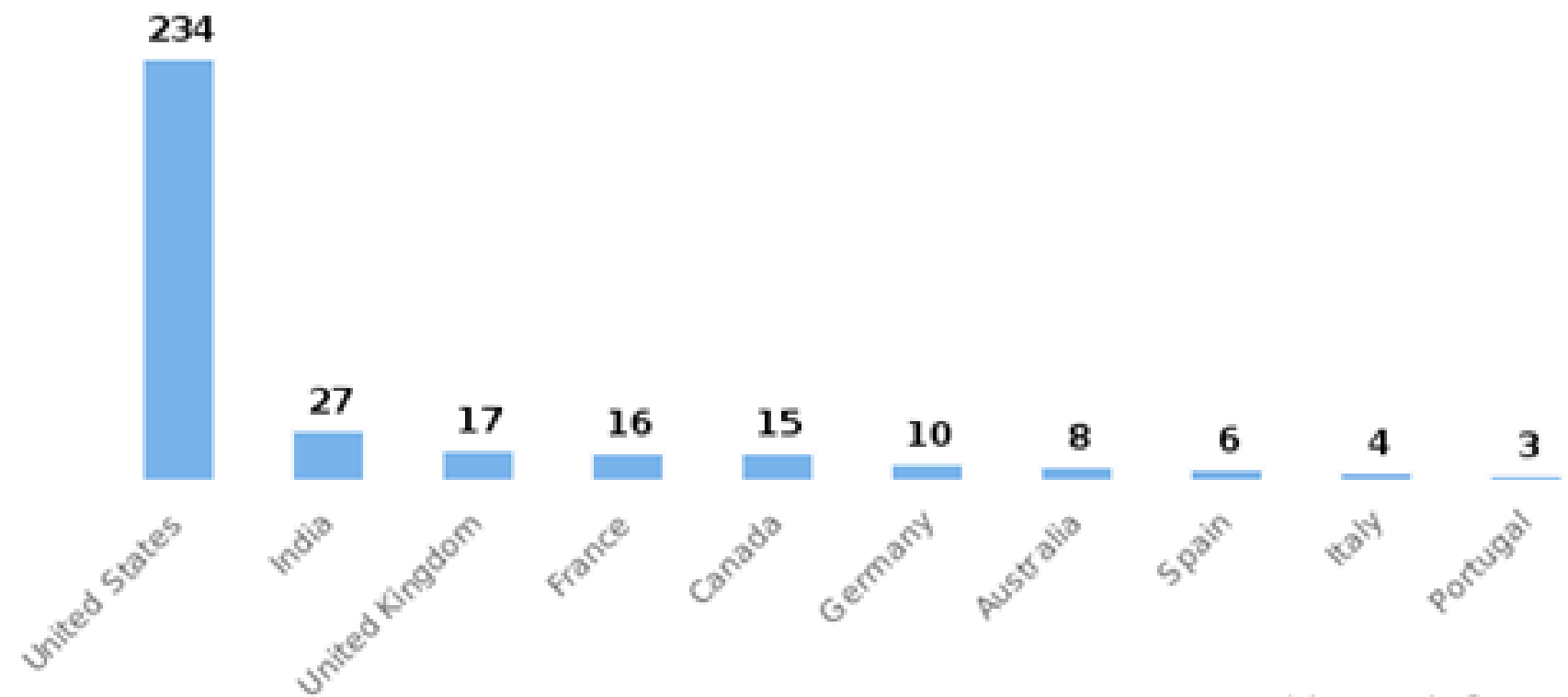
ỨNG DỤNG CỤ THỂ

Top Countries that use Apache Drill

Top Countries that use Apache Drill

60% of Apache Drill customers are in United States and 7% are in India.

Distribution of companies using Apache Drill by Country



powered by enlyft.com

ỨNG DỤNG CỤ THỂ



Báo cáo và vận hành BI



Phân tích việc sử dụng các thiết bị mạng không dây bằng cách chạy truy vấn SQL trên bộ dữ liệu lớn



Thực hiện ETL (trích xuất, chuyển đổi và tải).
Bao gồm Kafka, Amazon S3 và MongoDB cho kho dữ liệu

ỨNG DỤNG CỤ THỂ



Phân tích dữ liệu dịch vụ nhằm hiểu rõ hơn để đưa ra quyết định kinh doanh

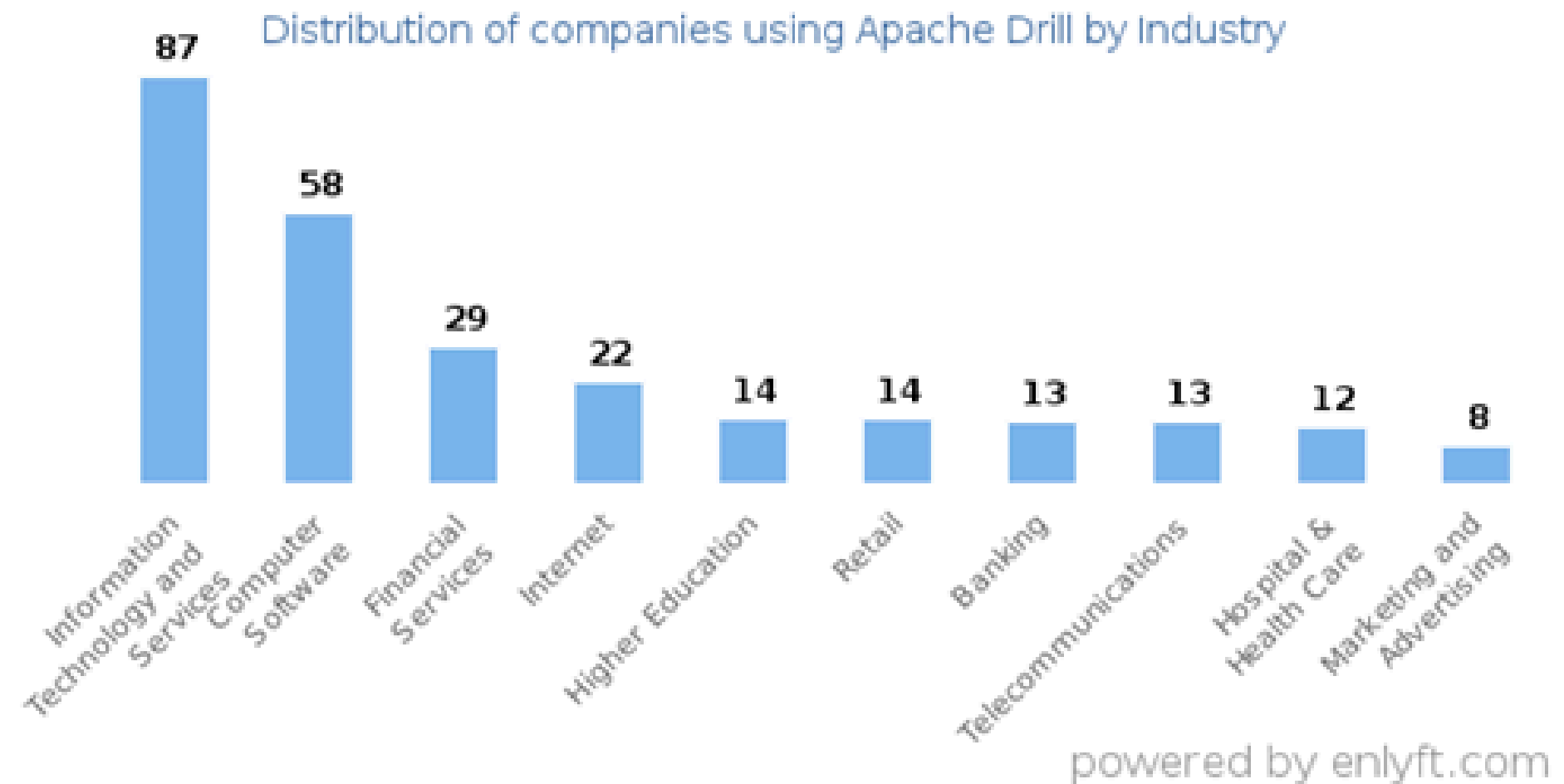


Phân tích và tạo báo cáo từ các chủ đề về luồng nhấp chuột của Kafka

ỨNG DỤNG CỤ THỂ

Top Industries that use Apache Drill

Looking at Apache Drill customers by industry, we find that Information Technology and Services (22%), Computer Software (15%), Financial Services (7%) and Internet (5%) are the largest segments.



TRIỂN KHAI APACHE DRILL

EMBEDDED MODE

```
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.2# ./bin/drill-embedded  
Apache Drill 1.21.2  
"Keep your data close, but your Drillbits closer."  
apache drill> █
```


EMBEDDED MODE

Apache Drill

localhost:8047

OptionsDocumentation

Drillbits1

#	Address ⓘ	Heap Memory Usage ⓘ	Direct Memory Usage ⓘ	CPU Usage ⓘ	Avg Sys Load ⓘ	User Port	Control Port	Data Port	Version	Status	Uptime	Shutdown
1	thuantv-VMware-Virtual-Platform Current	0.51GB (13% of 4GB)	0GB (0% of 0GB)	0.30%	0.20	31010	31011	31012	1.21.2	ONLINE	1m 51s	

Encryption

Client to Bit Encryption

Disabled

Bit to Bit Encryption

Disabled

Query Throttling

Queue Status

Disabled

EMBEDDED MODE

```
{
  // string, 22 character unique string business id
  "business_id": "tnhfDv5Il8EaGSXZGiuQGg",

  // string, the business's name
  "name": "Garaje",

  // string, the full address of the business
  "address": "475 3rd St",

  // string, the city
  "city": "San Francisco",

  // string, 2 character state code, if applicable
  "state": "CA",

  // string, the postal code
  "postal code": "94107",

  // float, latitude
  "latitude": 37.7817529521,

  // float, longitude
  "longitude": -122.39612197,

  // float, star rating, rounded to half-stars
  "stars": 4.5,

  // integer, number of reviews
  "review_count": 1198,
```

```
  // integer, 0 or 1 for closed or open, respectively
  "is_open": 1,

  // object, business attributes to values. note: some attribute values m
  "attributes": {
    "RestaurantsTakeOut": true,
    "BusinessParking": {
      "garage": false,
      "street": true,
      "validated": false,
      "lot": false,
      "valet": false
    },
  },

  // an array of strings of business categories
  "categories": [
    "Mexican",
    "Burgers",
    "Gastropubs"
  ],

  // an object of key day to value hours, hours are using a 24hr clock
  "hours": {
    "Monday": "10:00-21:00",
    "Tuesday": "10:00-21:00",
    "Friday": "10:00-21:00",
    "Wednesday": "10:00-21:00",
    "Thursday": "10:00-21:00",
    "Sunday": "11:00-18:00",
    "Saturday": "10:00-21:00"
  }
}
```

EMBEDDED MODE

Truy vấn 1: Tổng quan dữ liệu business

```
SELECT *
FROM dfs.`/home/thuantv/yelp_dataset/yelp_academic_dataset/business.json`LIMIT 15;
```

Query Profile: 18e60f5e-e143-3d17-7eac-3649fcd9c14c COMPLETED

Column visibility

Show

10

 entries

Delimiter

,

Export

Search:

business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	attributes
Pns2I4eNsIO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.4266787	-119.7111968	5.0	7	0	{"ByAppointmentC
mpf3x-BjTdTEA3yCZrAYPw	The UPS Store	87 Grasso Plaza Shopping Center	Afton	MO	63123	38.551126	-90.335695	3.0	15	1	{"BusinessAccept
tUFrWirKIKI_TAnsVWINQQ	Target	5255 E Broadway Blvd	Tucson	AZ	85711	32.223236	-110.880452	3.5	22	0	{"ByAppointmentC
MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.9555052	-75.1555641	4.0	80	1	{"ByAppointmentC
mWWMc6_wTdE0EUBKIGXDVIa	Perkiomen Valley Brewery	101 Walnut St	Green Lane	PA	18054	40.3381827	-75.4716585	4.5	13	1	{"BusinessAccept
CF33F8-E6oudUQ46HnavjQ	Sonic Drive-In	615 S Main St	Ashland City	TN	37015	36.269593	-87.058943	2.0	6	1	{"ByAppointmentC
n_0UpQx1hsNbnPUSlodU8w	Famous Footwear	8522 Eager Road, Dierbergs Brentwood Point	Brentwood	MO	63144	38.627695	-90.340465	2.5	13	1	{"BusinessAccept
qkRM_2X51Yqpk3b0wAQlg	Temple Beth-El	400 Pasadena Ave S	St. Petersburg	FL	33707	27.76659	-82.732983	3.5	5	1	{}
k0hBqXX-Bt0vf1op7Jr1w	Tsevi's Pub And Grill	8025 Mackenzie Rd	Afton	MO	63123	38.5651648	-90.3210868	3.0	19	0	{"BusinessAccept
bBDDEgkFA1Oxc9Lfe7BZUQ	Sonic Drive-In	2312 Dickerson Pike	Nashville	TN	37207	36.2081024	-86.7681696	1.5	10	1	{"ByAppointmentC

Showing 1 to 10 of 15 entries

Previous

1

2

Next

EMBEDDED MODE

Truy vấn 2: Tổng số lượt đánh giá của dataset

```
SELECT sum(review_count) as totalreviews
FROM dfs.`/home/thuantv/yelp_dataset/yelp_academic_dataset_business.json`;
```

←

→

↺

localhost:8047/query

☆

🔒

👤

🏠

☰

Apache Drill

Query

Profiles

Storage

Metrics

Threads

Logs

Options

Documentation

Query Profile: 18e60e17-d3fe-6122-eece-1ef3621b9aeb

COMPLETED

🔗

Delimiter

,

Export

⬆

Column visibility

Show

10

▼

entries

Search:

totalreviews
6745508

EMBEDDED MODE

Truy vấn 3: Top 10 các bang và thành phố có số review cao nhất

```
SELECT state, city, count(*) totalreviews
FROM dfs.`/home/thuante/yelp_dataset/yelp_academic_dataset_business.json`
GROUP BY state, city ORDER BY count(*) DESC LIMIT 10;
```

state	city	totalreviews
PA	Philadelphia	14567
AZ	Tucson	9249
FL	Tampa	9048
IN	Indianapolis	7540
TN	Nashville	6968
LA	New Orleans	6208
NV	Reno	5932
AB	Edmonton	5054
MO	Saint Louis	4827
CA	Santa Barbara	3829

EMBEDDED MODE

Truy vấn 4: Thời gian mở cửa, đóng cửa của các doanh nghiệp vào thứ 7

```
SELECT b.name, b.hours.Saturday`  
FROM dfs.`/home/thuante/yelp_dataset/yelp_academic_dataset_business.json`b LIMIT 10;
```

name	EXPR\$1
Abby Rappoport, LAC, CMQ	null
The UPS Store	8:0-14:0
Target	8:0-23:0
St Honore Pastries	7:0-21:0
Perkiomen Valley Brewery	12:0-22:0
Sonic Drive-In	9:0-22:0
Famous Footwear	10:0-18:0
Temple Beth-El	null
Tsevi's Pub And Grill	null
Sonic Drive-In	6:0-17:0

DISTRIBUTED MODE

Open  

zoo.cfg

~/apache-zookeeper-3.5.10-bin/conf

```
tickTime=2000
dataDir=/home/thuantv/apache-zookeeper-3.5.10-bin/data
clientPort=2182
initLimit=5
syncLimit=2
server.1=127.0.0.1:2888:3888
```

DISTRIBUTED MODE

```
drill.exec: {  
  cluster-id: "drill-cluster",  
  zk.connect: "127.0.0.1:2182",  
  drillbit: {  
    port: 31010,  
    http: {  
      port: 8047  
    }  
  },  
  exec: {  
    allow_loopback_address_binding: true  
  }  
}
```

DISTRIBUTED MODE

```
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-zookeeper-3.5.10-bin#  
./bin/zkServer.sh start  
ZooKeeper JMX enabled by default  
Using config: /home/thuantv/apache-zookeeper-3.5.10-bin/bin/../conf/zoo.cfg  
Starting zookeeper ... STARTED  
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-zookeeper-3.5.10-bin#  
./bin/zkServer.sh status  
ZooKeeper JMX enabled by default  
Using config: /home/thuantv/apache-zookeeper-3.5.10-bin/bin/../conf/zoo.cfg  
Client port found: 2182. Client address: localhost. Client SSL: false.  
Mode: standalone  
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-zookeeper-3.5.10-bin#
```

Start Zookeeper

DISTRIBUTED MODE

```
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.2# ./bin/drillbit.sh start  
Starting drillbit, logging to /home/thuantv/apache-drill-1.21.2/log/drillbit.out  
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.2# ./bin/drillbit.sh status  
drillbit is running.  
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.2# ./bin/drillbit.sh status  
drillbit is running.
```

Start Drill


DISTRIBUTED MODE

Apache Drill

QueryProfilesStorageMetricsThreadsLogs

OptionsDocumentation

Drillbits1

#	Address ⓘ	Heap Memory Usage ⓘ	Direct Memory Usage ⓘ	CPU Usage ⓘ	Avg Sys Load ⓘ	User Port	Control Port	Data Port	Version	Status	Uptime	Shutdown
1	thuantv-VMware-Virtual-Platform Current	0.22GB (5% of 4GB)	0GB (0% of 0GB)	0.27%	0.11	31010	31011	31012	1.21.2	ONLINE	3m 38s	

Encryption

Client to Bit Encryption

Disabled

Bit to Bit Encryption

Disabled

Query Throttling

Queue Status

Disabled

DISTRIBUTED MODE

```
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.1# ./bin/drillbit.sh start
Starting drillbit, logging to /home/thuantv/apache-drill-1.21.1/log/drillbit.out
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.1# ./bin/drillbit.sh status
drillbit is running.
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.1# ./bin/drillbit.sh status
drillbit is running.
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.1# ./bin/drillbit.sh status
drillbit is running.
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.1# ./bin/drillbit.sh status
drillbit is running.
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.1# ./bin/drillbit.sh status
drillbit is running.
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.1# ./bin/drillbit.sh status
/home/thuantv/apache-drill-1.21.1/drillbit.pid file is present but drillbit is not running.
root@thuantv-VMware-Virtual-Platform:/home/thuantv/apache-drill-1.21.1#
```

Lỗi drillbit.pid file is present but drillbit is not running

**CẢM ƠN THẦY VÀ CÁC BẠN
ĐÃ CHÚ Ý LẮNG NGHE**