Leveraging SAP HANA & Hortonworks Data Platform to analyze Wikipedia Page Hit Data

1 Introduction

SAP HANA is the leading OLTP and OLAP platform delivering instant access and critical business insight using a single in-memory columnar data store. Hortonworks Data Platform is the leading 100% open source, fully tested and certified Apache Hadoop platform enabling businesses to store, enrich and analyze the ever-increasing volumes of observational data such as web-clicks, sensor readings and social sentiments. Together SAP HANA and Hortonworks Data platform (HDP) offer unique opportunity for enterprises to cost effectively capture all enterprise data (ERP & non-ERP) and analyze them in real time to drive efficiency and enable new business opportunities.

This demonstration showcases a common integration scenario of SAP HANA and HDP, where-in raw observational data (web-clicks) is collected in Hortonworks Data Platform. The raw data is enriched and filtered in Hortonworks Data Platform to create data subsets critical for further analysis. Filtered dataset is subsequently transferred to HANA for real-time aggregation and predictive analytics.

2 Audience

The demonstration is targeted for solution architects and field engineers interested in understanding and showcasing the joint capabilities of SAP HANA and Hortonworks Data Platform.

Demonstration setup instructions leverage prior work and refer to them where appropriate.

3 Prerequisites

To understand and enable a working demonstration, following conditions must be met:

1. User must have basic knowledge of Linux
2. User must have an Amazon Web Services (AWS) account
    - Users sign up for an AWS account at http://aws.amazon.com
3. User must have a basic understanding of AWS

4 Cost

The demonstration leverages SAP HANA Developer edition and Hortonworks Data Platform, both available at no software charge. However, the user is responsible for the Amazon EC2 instance charges that serve as the hardware infrastructure for the demonstration scenario. Based on recommended hardware configurations, the following charges are anticipated:

Server	Specification	Projected Cost/Per Hour
EC2 server for SAP HANA Developer Edition	EBS Optimized M1.xlarge 250 GB EBS Block Storage	$0.65*
EC2 server for Hortonworks Data Platform	M2.xlarge 130 GB EBS Block Storage	0.550*
Total Cost		$1.20/per hour

* Cost per hour is an approximation based on EC2 rates as of December 2013. Above server specifications are recommended for balanced cost and performance.

5 Demonstration Scenario

The demonstration leverages HDP and HANA to analyze Wikipedia web click data. Wikipedia maintains webclick statistics for it's pages and makes them available for download as compressed files. The demonstration illustrates the following scenario:

1) Download Wikipedia pagehit data and load it to HDP
2) Leverage HDP to filter raw data, retaining pages with hit count of over 99
    - Filtering focuses the analysis on high value pages
3) Retrieve filtered data in HDP into SAP HANA using Smart Data Access
4) Develop analytical view and visualize the data using SAP Hana Studio

6 Setup Instructions

Following steps must be followed to setup the demonstration:

1) Setup SAP HANA, developer edition on Amazon Web Services (AWS) and HANA Studio
    a. Note: SAP HANA on AWS is not mandatory -- if you have access to your own instances of SAP HANA, you can leverage your current instance.

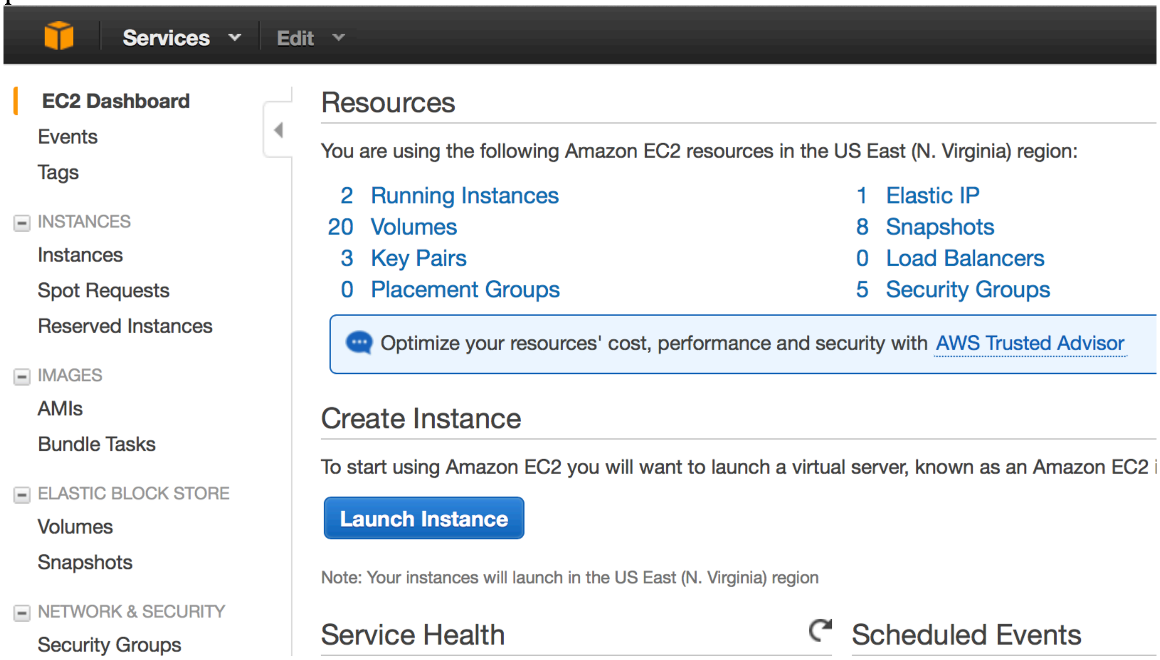
2) Setup Hortonworks Data Platform on AWS
3) Install Hortonworks Hive ODBC Driver on SAP HANA AWS server
4) Access HDP from SAP HANA using Smart Data Access

6.1 Setup SAP HANA Developer Edition on AWS & SAP HANA Studio

Detailed step-by-step instruction to create SAP HANA Developer Edition on AWS and setting HANA Studio has been published by SAP. Follow the instructions as outlined in the document http://scn.sap.com/docs/DOC-28294.

6.2 Setup Hortonworks Data Platform on AWS

- Log into your AWS console and click on the AMI menu in the left navigation pane



Resources

You are using the following Amazon EC2 resources in the US East (N. Virginia) region:

2	Running Instances	1	Elastic IP
20	Volumes	8	Snapshots
3	Key Pairs	0	Load Balancers
0	Placement Groups	5	Security Groups

Optimize your resources' cost, performance and security with AWS Trusted Advisor

Create Instance

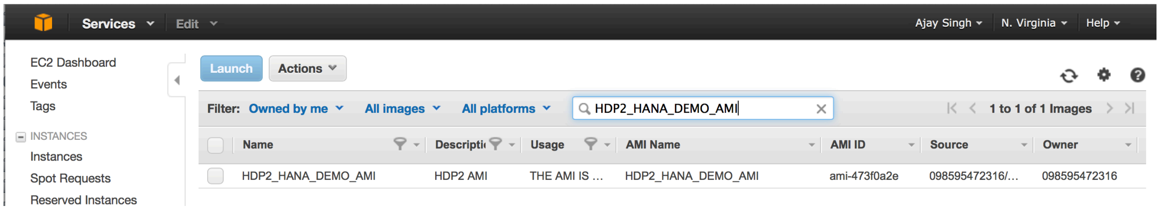
To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 i

Note: Your instances will launch in the US East (N. Virginia) region

Service Health

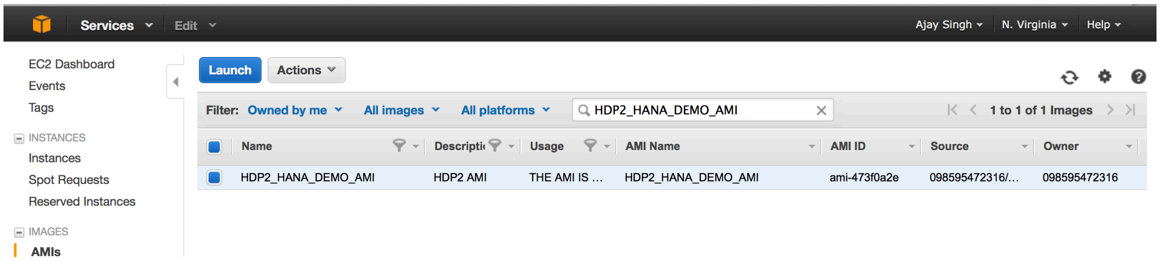
Scheduled Events

- Search for the AMI “HDP2_HANA_DEMO_AMI” in the **US East Region**



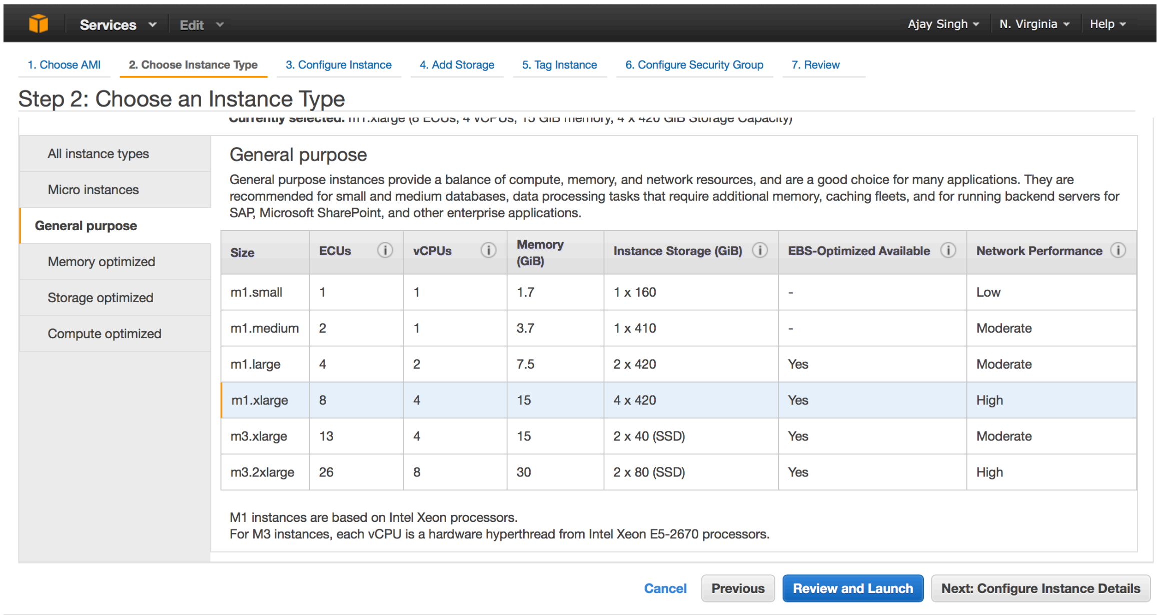
Name	Descripti	Usage	AMI Name	AMI ID	Source	Owner
HDP2_HANA_DEMO_AMI	HDP2 AMI	THE AMI IS ...	HDP2_HANA_DEMO_AMI	ami-473f0a2e	098595472316/...	098595472316

- Select the HDP2_HANA_DEMO_AMI and start it using the Launch button at the top.



☑	Name	Descripti	Usage	AMI Name	AMI ID	Source	Owner
☑	HDP2_HANA_DEMO_AMI	HDP2 AMI	THE AMI IS ...	HDP2_HANA_DEMO_AMI	ami-473f0a2e	098595472316/...	098595472316

- In the resulting wizard, select m1.xlarge for the Instance Type by selecting General purpose menu on the left. Then, click “Next Configure Instance Details”.



Step 2: Choose an Instance Type

Currently selected: m1.xlarge (8 ECUs, 4 vCPUs, 15 GiB memory, 4 x 420 GiB Storage Capacity)

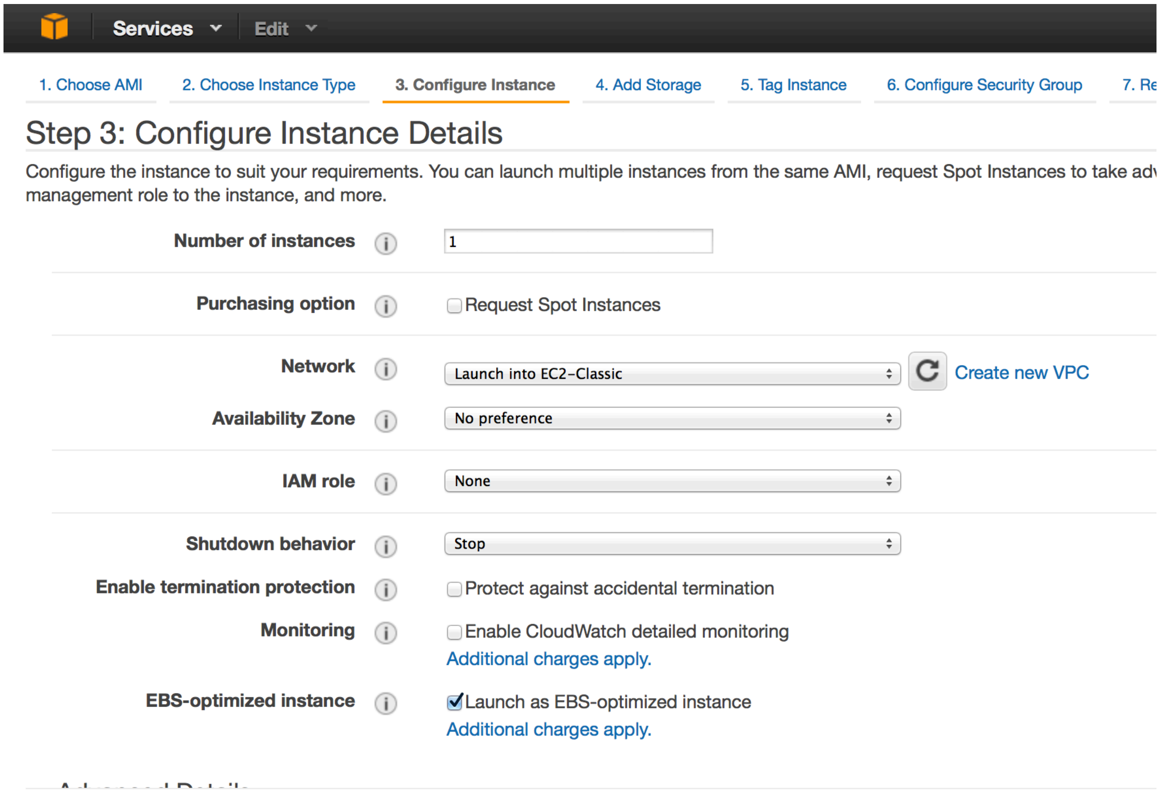
General purpose

General purpose instances provide a balance of compute, memory, and network resources, and are a good choice for many applications. They are recommended for small and medium databases, data processing tasks that require additional memory, caching fleets, and for running backend servers for SAP, Microsoft SharePoint, and other enterprise applications.

Size	ECUs	vCPUs	Memory (GiB)	Instance Storage (GiB)	EBS-Optimized Available	Network Performance
m1.small	1	1	1.7	1 x 160	-	Low
m1.medium	2	1	3.7	1 x 410	-	Moderate
m1.large	4	2	7.5	2 x 420	Yes	Moderate
m1.xlarge	8	4	15	4 x 420	Yes	High
m3.xlarge	13	4	15	2 x 40 (SSD)	Yes	Moderate
m3.2xlarge	26	8	30	2 x 80 (SSD)	Yes	High

M1 instances are based on Intel Xeon processors.
For M3 instances, each vCPU is a hardware hyperthread from Intel Xeon E5-2670 processors.

- In the step 3 “Configure Instance Details”, select EBS-optimized instance and leave the remaining values as their defaults. Click Next:Add Storage when done.



Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot Instances to take adv management role to the instance, and more.

Number of instances 1

Purchasing option □ Request Spot Instances

Network Launch into EC2-Classic Create new VPC

Availability Zone No preference

IAM role None

Shutdown behavior Stop

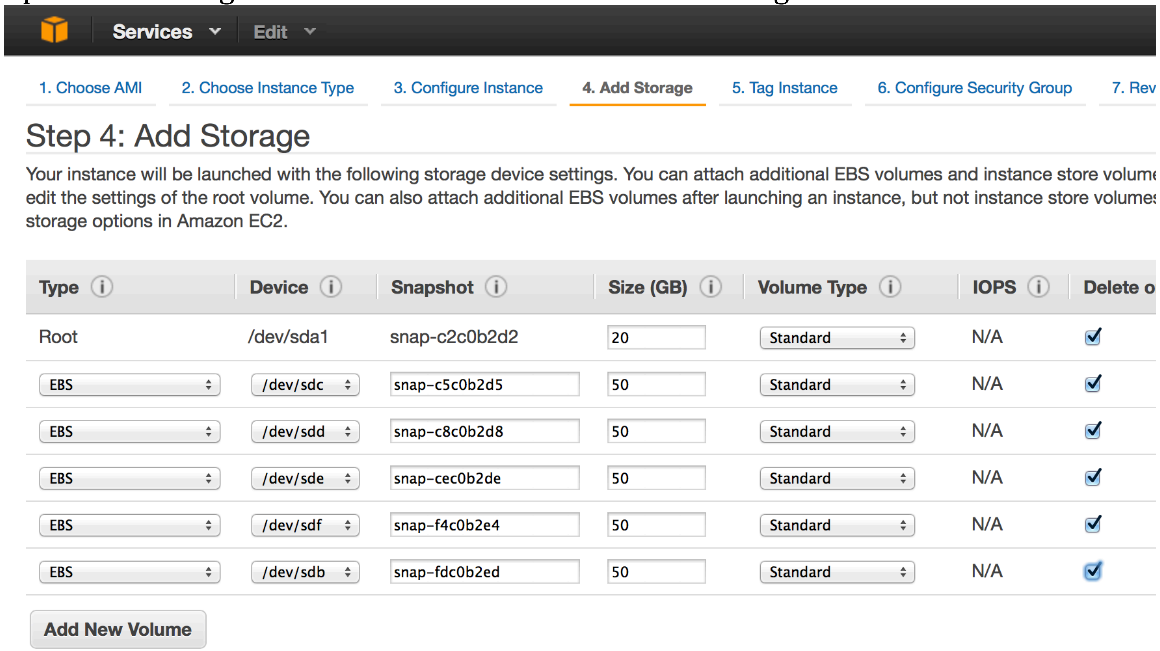
Enable termination protection □ Protect against accidental termination

Monitoring □ Enable CloudWatch detailed monitoring
Additional charges apply.

EBS-optimized instance ☑ Launch as EBS-optimized instance
Additional charges apply.

Advanced Details

- Step 4 “Add Storage” leave the defaults and select Next: Tag Instance

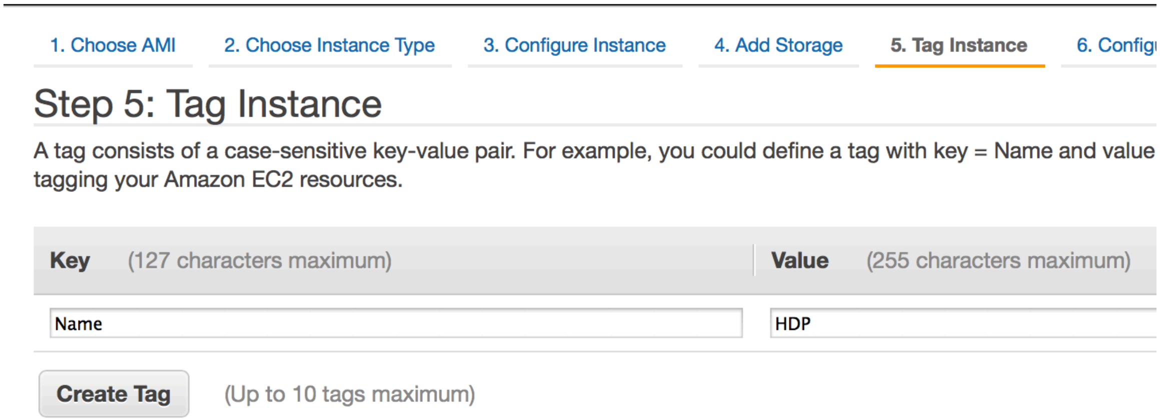


Step 4: Add Storage

Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volume edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes storage options in Amazon EC2.

Type	Device	Snapshot	Size (GB)	Volume Type	IOPS	Delete o
Root	/dev/sda1	snap-c2c0b2d2	20	Standard	N/A	☑
EBS	/dev/sdc	snap-c5c0b2d5	50	Standard	N/A	☑
EBS	/dev/sdd	snap-c8c0b2d8	50	Standard	N/A	☑
EBS	/dev/sde	snap-cec0b2de	50	Standard	N/A	☑
EBS	/dev/sdf	snap-f4c0b2e4	50	Standard	N/A	☑
EBS	/dev/sdb	snap-fdc0b2ed	50	Standard	N/A	☑

- Step 5 “Tag Instance”, add the tag “Name = HDP”. Click Next: Configure Security Group when done.



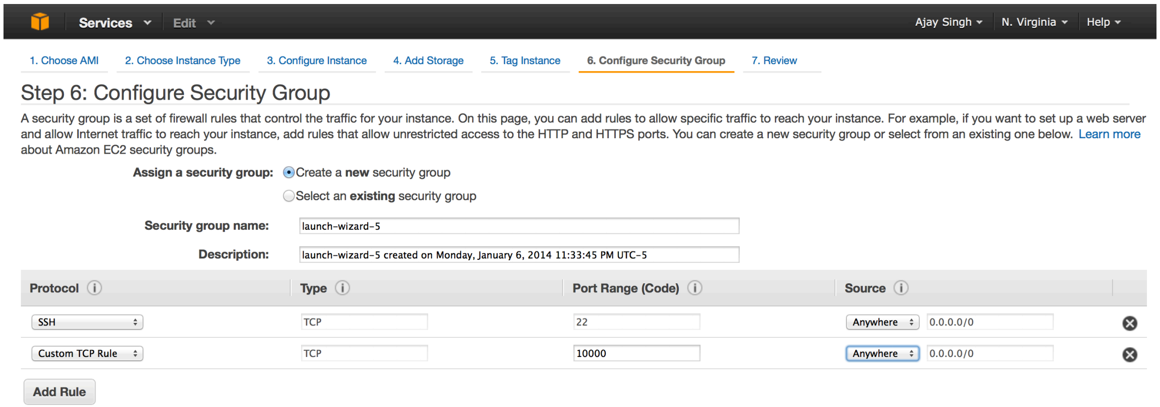
Step 5: Tag Instance

A tag consists of a case-sensitive key-value pair. For example, you could define a tag with key = Name and value tagging your Amazon EC2 resources.

Key (127 characters maximum)	Value (255 characters maximum)
Name	HDP

- ➢ Step 6 “Configure Security Group”, select “Create a new security group” with the following rules

Protocol	Type	Port Range	Source
• SSH	• TCP	• 22	• Anywhere
• Custom TCP Rule	• TCP	• 10000	• Anywhere



Step 6: Configure Security Group

A security group is a set of firewall rules that control the traffic for your instance. On this page, you can add rules to allow specific traffic to reach your instance. For example, if you want to set up a web server and allow Internet traffic to reach your instance, add rules that allow unrestricted access to the HTTP and HTTPS ports. You can create a new security group or select from an existing one below. Learn more about Amazon EC2 security groups.

Assign a security group: ● Create a **new** security group
○ Select an **existing** security group

Security group name: launch-wizard-5

Description: launch-wizard-5 created on Monday, January 6, 2014 11:33:45 PM UTC-5

Protocol	Type	Port Range (Code)	Source
SSH	TCP	22	Anywhere 0.0.0.0/0
Custom TCP Rule	TCP	10000	Anywhere 0.0.0.0/0

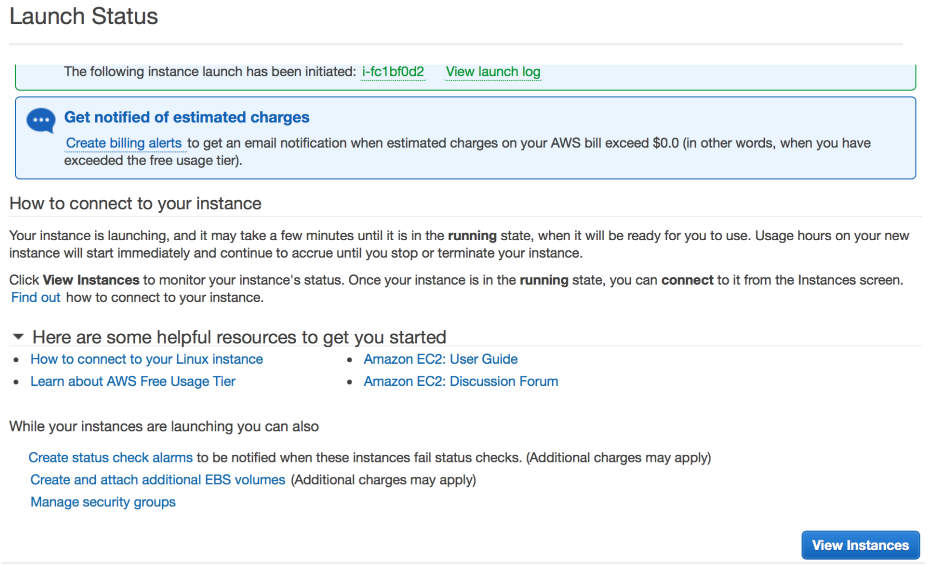
Once done select “Review and Launch”

- In the review window, validate the following
    - Instance Type matches the settings below:
        - Instance Type: m1.xlarge
        - EBS-Optimized: Yes
        - Network Performance: High
    - Security Groups matches the settings below

Protocol	Type	Port Range	Source
• SSH	• TCP	• 22	• 0.0.0.0/0
• Custom TCP Rule	• TCP	• 10000	• 0.0.0.0/0

    - Click Launch when done

- Select a key in the resulting window and launch the instance.



Launch Status

The following instance launch has been initiated: i-fc1bf0d2 View launch log

Get notified of estimated charges

Create billing alerts to get an email notification when estimated charges on your AWS bill exceed $0.0 (in other words, when you have exceeded the free usage tier).

How to connect to your instance

Your instance is launching, and it may take a few minutes until it is in the **running** state, when it will be ready for you to use. Usage hours on your new instance will start immediately and continue to accrue until you stop or terminate your instance.

Click **View Instances** to monitor your instance's status. Once your instance is in the **running** state, you can **connect** to it from the Instances screen. Find out how to connect to your instance.

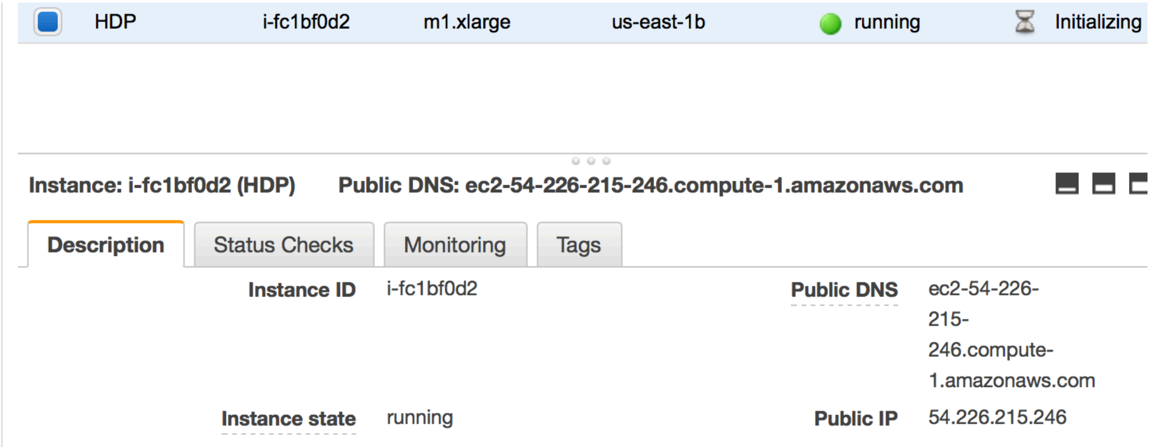
Here are some helpful resources to get you started

- How to connect to your Linux instance
- Learn about AWS Free Usage Tier
- Amazon EC2: User Guide
- Amazon EC2: Discussion Forum

While your instances are launching you can also

Create status check alarms to be notified when these instances fail status checks. (Additional charges may apply)
Create and attach additional EBS volumes (Additional charges may apply)
Manage security groups

- Navigate to the Instances view and note the public DNS & IP of the launched instance. We will use the public DNS & IP to establish a connection from SAP HANA.



HDP	i-fc1bf0d2	m1.xlarge	us-east-1b	running	Initializing

Instance: i-fc1bf0d2 (HDP) Public DNS: ec2-54-226-215-246.compute-1.amazonaws.com

Instance ID	i-fc1bf0d2	Public DNS	ec2-54-226-215-246.compute-1.amazonaws.com
Instance state	running	Public IP	54.226.215.246

The instance will take a few minutes to initialize. Upon initialization, the instance will contain a running instance of Hortonworks Data Platform with following key services enabled:

- HDFS
- YARN
- MapReduce HistoryServer
- Amabri
- Hive
- Zookeeper

Sample Data Set

The Hortonworks Data Platform comes pre-loaded with Wikipedia page hit data for the Months of March and April, 2014. Following filtered datasets are available as hive tables.

Table Name	Description	Num of Records
pagecountfilter100	Pages that have received over 99 hits in an hour	22 Million
pagecountfilter1000	Pages that have received over 999 hits in an hour	9 Million
pagecountfilter10000	Pages that have received over 9999 hits in an hour	1 Million

Different size datasets have been made available to facilitate experimentation using variably sized SAP HANA & HDP AWS server instances. Note the demonstration leverages modest sized servers for SAP HANA and HDP. As such the usage of table pagecountfilter10000 is recommended.

All the tables share a common schema, as noted below:

Column Name	Column Type
project code	String
Pagename	String
Hitsperhour	Bigint
Bytesperhour	Bigint
Year	String
Month	String
Day	String
Hour	String

6.3 Install Hortonworks Hive ODBC Driver on SAP HANA AWS Instance

Now that we have successfully launched the SAP HANA Developer and HDP instances on AWS, we will next install the Hortonworks Hive ODBC driver on SAP HANA instance, so one can access data in Hortonworks Data Platform from within SAP HANA. To install the ODBC driver, perform the following steps:

Setup

- Log in SAP HANA Developer Edition as root
- In the shell window, run the following commands:

```
mkdir /tmp/HDP
cd /tmp/HDP
wget http://public-repo-1.hortonworks.com/HDP/hive-odbc/1.2.13.1018/suse11/hive-odbc-native-1.2.13.1018.tar.gz
tar -xvzf hive-odbc-native-1.2.13.1018.tar.gz
rpm -i hive-odbc-native-1.2.13.1018/hive-odbc-native-1.2.13.1018-1.x86_64.rpm
```

create file /usr/sap/HDB/home/.hortonworks.hiveodbc.ini with following content

```
[Driver]
DriverManagerEncoding=UTF-16
ErrorMessagesPath=/usr/lib/hive/lib/native/hiveodbc/ErrorMessages/
LogLevel=0
LogPath=
ODBCInstLib=libodbcinst.so
```

create the file /usr/sap/HDB/home/.odbc.ini and copy the contents below. Note that the HOST & HS2HostFQDN (identified in red) should be updated to the public DNS of Hortonworks Data Platform launched in step 6.2

```
[ODBC]
# Specify any global ODBC configuration here such as ODBC tracing.

[ODBC Data Sources]
HDP=Hortonworks Hive ODBC Driver 64-bit
[HDP]

Description=Hortonworks Hive ODBC Driver (64-bit) DSN
Driver=/usr/lib/hive/lib/native/Linux-amd64-64/libhortonworkshiveodbc64.so
HOST= ec2-54-225-6-165.compute-1.amazonaws.com
HS2HostFQDN= ec2-54-225-6-165.compute-1.amazonaws.com
PORT=10000
Schema=default
FastSQLPrepare=0
UseNativeQuery=0
HiveServerType=2
HS2AuthMech=2
UserName=hive
```

```
chmod 777 /usr/sap/HDB/home/.hortonworks.hiveodbc.ini
chmod 777 /usr/sap/HDB/home/.odbc.ini
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/lib/hive/lib/native/Linux-amd64-64
echo export LD_LIBRARY_PATH=\$LD_LIBRARY_PATH:/usr/lib/hive/lib/native/Linux-amd64-64 >> /usr/sap/HDB/home/.sapenv.sh
echo export
LD_LIBRARY_PATH=\$LD_LIBRARY_PATH:/usr/lib/hive/lib/native/Linux-
amd64-64 >> /usr/sap/HDB/home/.sapenv.csh
export ODBCINI=/usr/sap/HDB/home/.odbc.ini
echo export ODBCINI=/usr/sap/HDB/home/.odbc.ini >>
/usr/sap/HDB/home/.sapenv.sh
echo export ODBCINI=/usr/sap/HDB/home/.odbc.ini >>
/usr/sap/HDB/home/.sapenv.csh
su - hdbadm
HDB stop
HDB start
```

Test the Setup

Run the following command to validate the setup

```
isql –v HDP
```

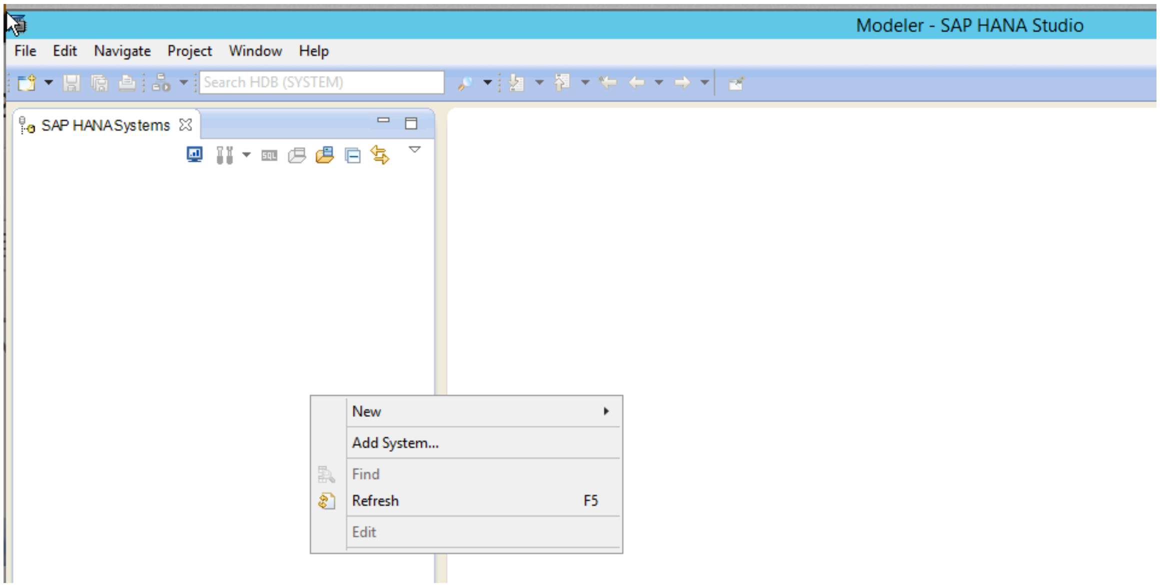
isql should successfully connect and show the sql prompt. In the event a connection is not established, please review to make sure all the shell commands above were executed successfully.

6.4 Access HDP from SAP HANA using Smart Data Access

As of this point, we have a working SAP HANA instance with Hortonworks Hive ODBC driver installed and a working Hortonworks Data Platform with sample data. We will now use SAP HANA Studio to establish a connection between HANA and HDP.

Establish connection between SAP Studio & SAP HANA on AWS

- Log into SAP HANA Studio and open the Modeler Perspective
- Add a new system

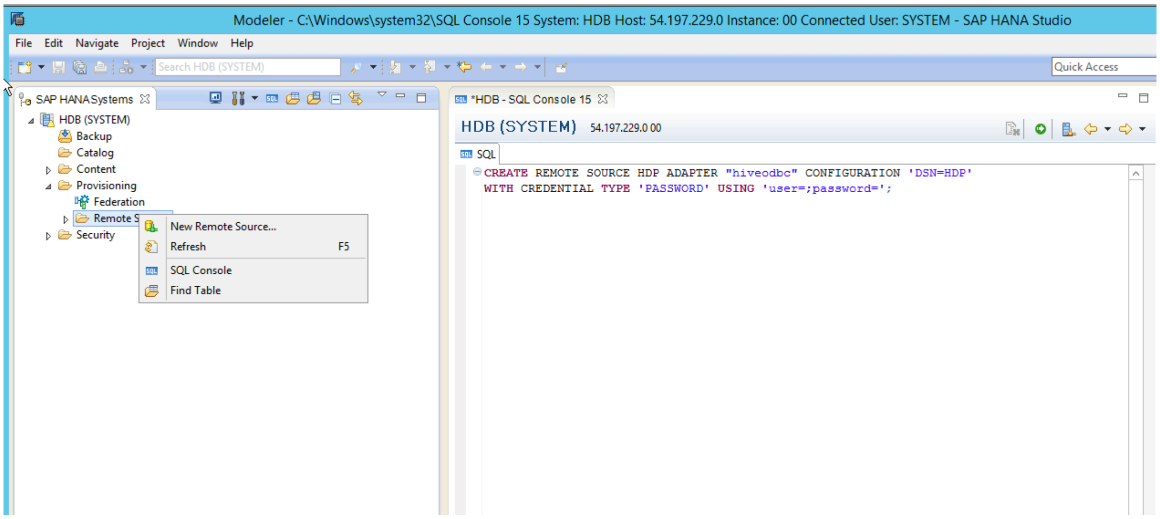


- In the resulting wizard, provide the following information:
  - Host Name: Public IP of your SAP HANA on AWS
  - Instance Number: 00
  - User Name: SYSTEM
  - Password: manager

Accept the default value for all the other fields and click Finish.

Create Virtual Table to access HDP from SAP HANA

- Navigate to Provisioning -> Remote Source and right click to select SQL Console


Modeler - C:\Windows\system32\SQL Console 15 System: HDB Host: 54.197.229.0 Instance: 00 Connected User: SYSTEM - SAP HANA Studio
File Edit Navigate Project Window Help
Search HDB (SYSTEM)
Quick Access
SAP HANASystems
HDB (SYSTEM)
Backup
Catalog
Content
Provisioning
Federation
Remote S
Security
New Remote Source...
Refresh F5
SQL Console
Find Table
*HDB - SQL Console 15
HDB (SYSTEM) 54.197.229.0 00
SQL

```
CREATE REMOTE SOURCE HDP ADAPTER "hiveodbc" CONFIGURATION 'DSN=HDP'
WITH CREDENTIAL TYPE 'PASSWORD' USING 'user=;password=';
```

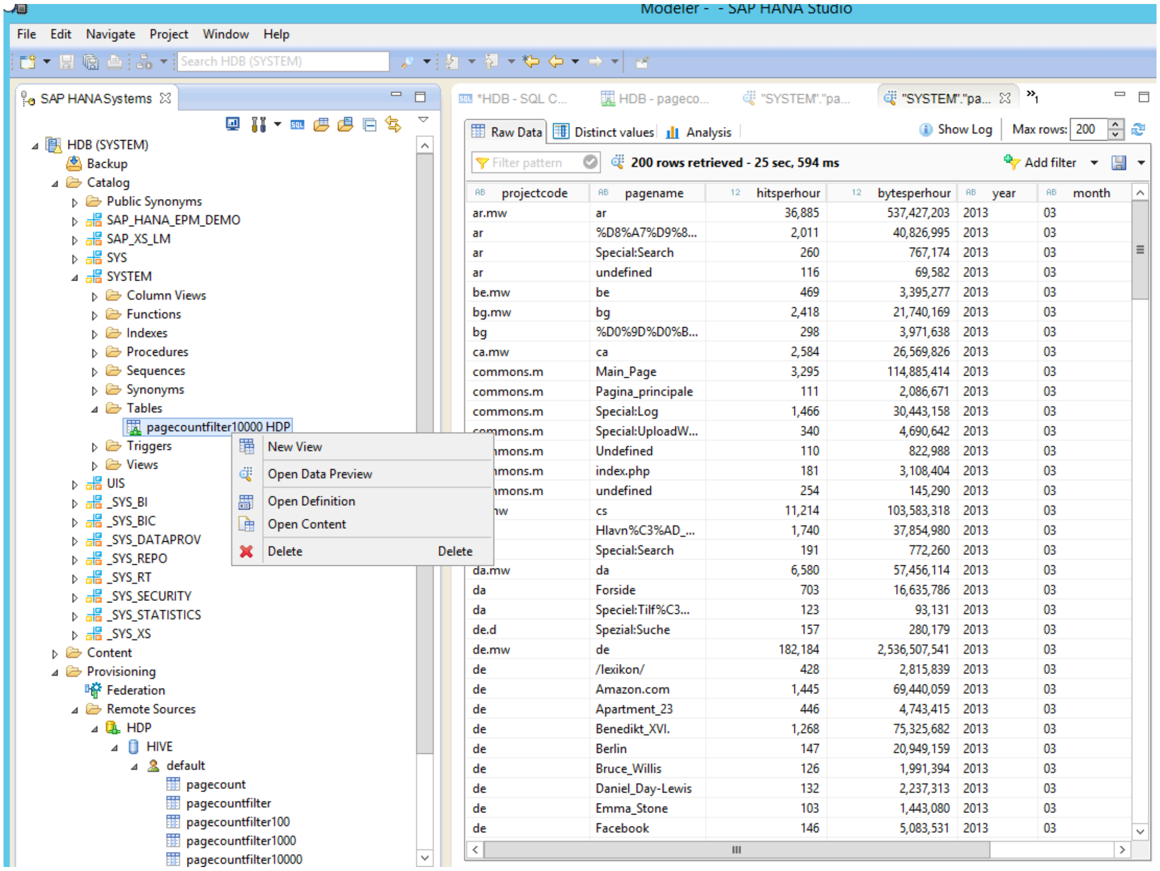
- In the SQL console, enter the following statement to create a remote data source

```
CREATE REMOTE SOURCE HDP ADAPTER "hiveodbc"
CONFIGURATION 'DSN=HDP' WITH CREDENTIAL TYPE
'PASSWORD' USING 'user=;password=';
```

- Create virtual table for HDP table pagehitfilter10000 using the following SQL command:

```
create virtual table "pagecountfilter10000"
("projectcode" varchar(50),
 "pagename" varchar(2000),
 "hitsperhour" BIGINT,
 "bytesperhour" BIGINT,
 "year" VARCHAR(4),
 "month" VARCHAR(2),
 "day" VARCHAR(2),
 "hour" VARCHAR(2)
 ) AT "HDP"."default"."default"."pagecountfilter10000";
```

- Review the data in the virtual table by selecting the table and opening it for Data preview



200 rows retrieved - 25 sec, 594 ms

projectcode	pagename	hitsperhour	bytesperhour	year	month
ar.mw	ar	36,885	537,427,203	2013	03
ar	%D8%A7%D9%8...	2,011	40,826,995	2013	03
ar	Special:Search	260	767,174	2013	03
ar	undefined	116	69,582	2013	03
be.mw	be	469	3,395,277	2013	03
bg.mw	bg	2,418	21,740,169	2013	03
bg	%D0%9D%D0%B...	298	3,971,638	2013	03
ca.mw	ca	2,584	26,569,826	2013	03
commons.m	Main_Page	3,295	114,885,414	2013	03
commons.m	Pagina_principale	111	2,086,671	2013	03
commons.m	Special:Log	1,466	30,443,158	2013	03
commons.m	Special:UploadW...	340	4,690,642	2013	03
mmons.m	Undefined	110	822,988	2013	03
mmons.m	index.php	181	3,108,404	2013	03
mmons.m	undefined	254	145,290	2013	03
nw	cs	11,214	103,583,318	2013	03
	Hlavn%C3%AD_...	1,740	37,854,980	2013	03
	Special:Search	191	772,260	2013	03
da.mw	da	6,580	57,456,114	2013	03
da	Forside	703	16,635,786	2013	03
da	Speciel:Tilf%C3...	123	93,131	2013	03
de.d	Spezial:Suche	157	280,179	2013	03
de.mw	de	182,184	2,536,507,541	2013	03
de	/lexikon/	428	2,815,839	2013	03
de	Amazon.com	1,445	69,440,059	2013	03
de	Apartment_23	446	4,743,415	2013	03
de	Benedikt_XVI.	1,268	75,325,682	2013	03
de	Berlin	147	20,949,159	2013	03
de	Bruce_Willis	126	1,991,394	2013	03
de	Daniel_Day-Lewis	132	2,237,313	2013	03
de	Emma_Stone	103	1,443,080	2013	03
de	Facebook	146	5,083,531	2013	03

Transfer data to SAP HANA for Instant Access and Interactive Analysis

- Create a column table using the following SQL

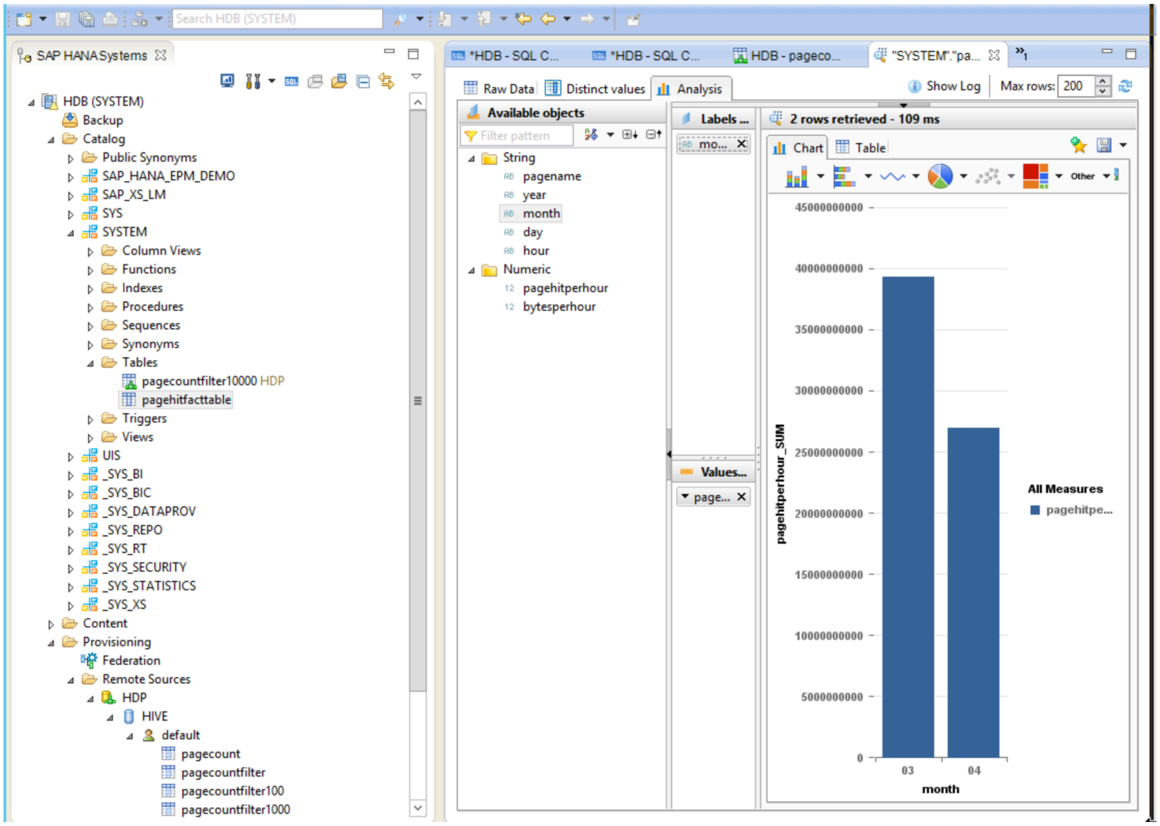
```
create column table "pagehitfacttable"
("pagename" varchar(2000),
 "year" varchar(4),
 "month" varchar(2),
 "day" varchar(2),
 "hour" varchar(2),
 "pagehitperhour" BIGINT,
"bytesperhour" BIGINT);
```

- Transfer data from virtual table pagecountfilter10000 to column table pagehitfacttable using the SQL

```
select
"pagename","year","month","day","hour","hitsperhour","bytesper
hour"
from "SYSTEM"."pagecountfilter10000"
where "hitsperhour" > 100000
into "SYSTEM"."pagehitfacttable";
```

Analyze the Column Table in SAP HANA

Select the table pagehitfacttable and open it in Data Preview mode for further analysis.



Max rows: 200

2 rows retrieved - 109 ms

7 Summary

The demonstration illustrates the unique capability of SAP HANA and Hortonworks Data Platform to deliver instant access at infinite scale. Organizations can leverage the joint capabilities of HANA and HDP to meet the needs of ever growing data volumes and a business imperative to act in real-time.

8 References

A number of blog posts were referenced to create the demonstration scenario. Of special note is the blog post by Bill Ramos on Importing Wikipedia Hive data into SAP HANA One. The original work by Bill was extended to include a live connection between Hortonworks Data Platform and SAP HANA.

9 Questions

If you have any question please reach us at sap@hortonworks.com