

SampleApp BigData Examples

- Contact your local Oracle Sales rep for details on enabling BigData on your SampleAppv607 virtual box image
- See examples below

Big Data Related Samples



What's New In CDH 5.7.x

Continue reading:

- What's New in CDH 5.7.0
- What's New in CDH 5.7.1

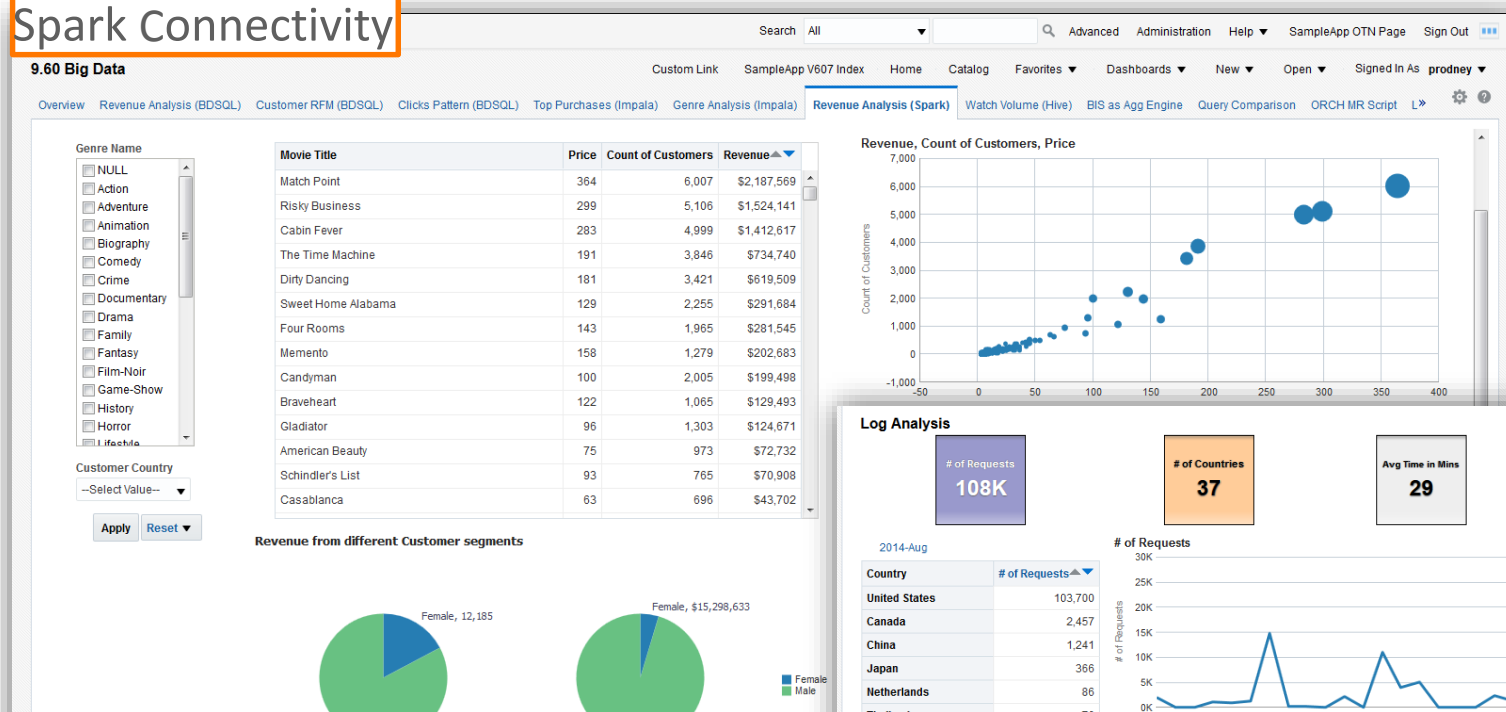
What's New in CDH 5.7.0

The following sections describe new features introduced in CDH

- Operating System Support
- Apache Hadoop
- Apache HBase
- Apache Hive
- Hue
- Apache Impala (incubating)
- MapReduce
- Apache Oozie
- Cloudera Search
- Apache Spark
- Apache Sentry
- YARN

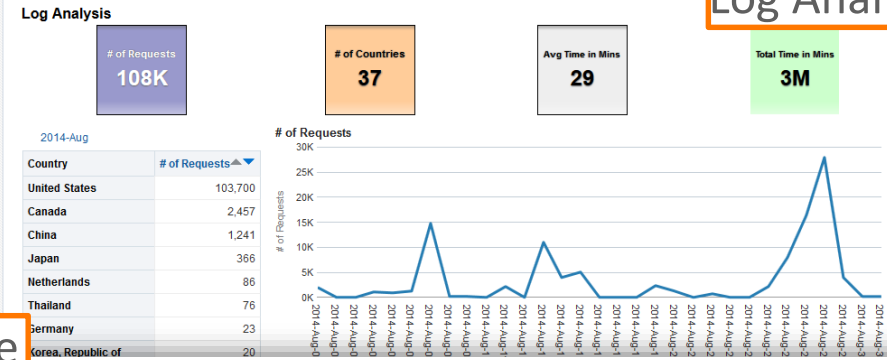


Spark Connectivity



CDH 5.7, Spark, BDSQL

Log Analysis



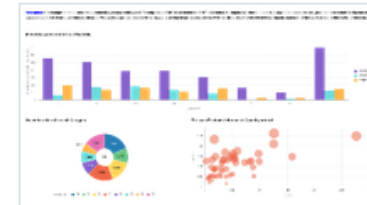
DV Projects – Spark, Impala, BDSQL, Hive



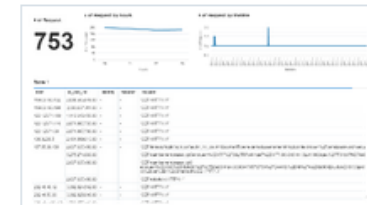
Spark-Online Sales Analysis
Visual Analyzer 3:57 PM



Hive - Customer Sales Analysis
Visual Analyzer 3:55 PM



Impala-Inventory Analysis
Visual Analyzer 3:56 PM



Hive - Log Analysis
Visual Analyzer 4:54 PM



What's New in SampleApp V606 - Big Data Components

Cloudera Hadoop 5.7

SampleApp 606 ships with CDH 5.7. CDH 5.7 has lot of new upgrades and improvements on Apache Hive, Hue, YARN and Spark etc, which in short means,now users can read data of even more data types and can store & retrieve data on an even massive scale, with enhanced security. For detailed information on CDH 5.7 please refer to the following link:

http://www.cloudera.com/documentation/enterprise/release-notes/topics/cdh_rm_new_in_cdh_57.html

Spark

With CDH 5.7 Spark is configured on YARN which helps in performing memory management efficiently ensuring that Spark is not starved for memory. Spark-SQL also comes pre-configured with CDH 5.7 and allows the user to directly invoke Spark-SQL without having to do any additional configurations

Spark Thrift Server

CDH 5.7 ships with Spark Thrift Server which is configured on YARN and it allows JDBC, ODBC connections to Spark Database. This means that now users can connect to Spark Database from various Oracle Analytic tools like Data Visualization Desktop, BICS, DVCS & OBIEE

Updated BDSQL

With the latest April-2016 Database Bundle Patch on the underlying Oracle Database. BDSQL is enabled on Oracle Database in SampleApp and it now supports CDH 5.7.

Visit this Dashboard for complete content

9.60 Big Data

Overview Revenue Analysis (BDSQL) Customer RFM (BDSQL) Clicks Pattern (BDSQL) Top Purchases (Impala) Genre Analysis (Impala)

Revenue Analysis (Spark) Watch Volume (Hive) BIS as Agg Engine Query Comparison ORCH MR Script

Page Information (click to collapse or expand)

Description : This report analyse revenues and number of distinct customers for each movie genre. Oracle Big Data SQL is used to read application activity logs residing in HDFS, which contains the movie purchase details. That is joined with Customer and Movie attributes data in Oracle database to present this report.

Month

- (All Column Values)
- 2007-07
- 2007-08
- 2007-09
- 2007-10

Customer Country

--Select Value--

Customer City

--Select Value--

Income Level

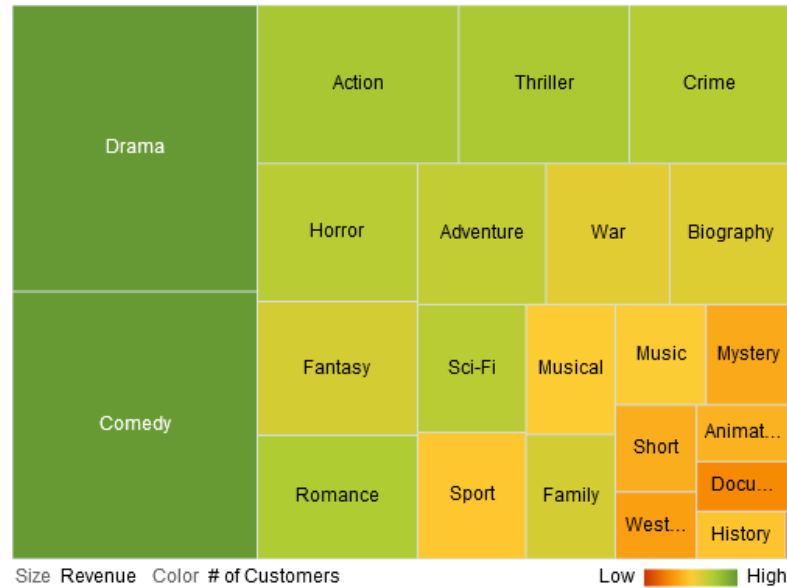
--Select Value--

Movie Genre

- NULL
- Action
- Adventure
- Animation
- Biography

Revenue Analysis

Genre Name	Revenue	# of Customers
Drama	\$1,602	1,244
Comedy	\$1,497	1,235
Action	\$728	962
Thriller	\$616	951
Crime	\$583	915
Horror	\$505	896
Fantasy	\$489	797
Romance	\$453	935
Adventure	\$413	864
War	\$398	741
Biography	\$389	755
Sci-Fi	\$317	892
Sport	\$314	605
Musical	\$265	631
Family	\$254	795



Description : This demonstrates Spark connectivity to data residing in HDFS using Spark ODBC driver included in OBIEE installation. The Spark SQL dialect supports a range of standard elements, plus some extensions for Big Data use cases related to data loading and data warehousing. Spark provides a high degree of compatibility with the Hive Query Language (HiveQL). Refer Spark SQL Language Reference documentation from Cloudera for more details.

Genre Name

- NULL
- Action
- Adventure
- Animation
- Biography
- Comedy
- Crime
- Documentary
- Drama
- Family
- Fantasy
- Film-Noir
- Game-Show
- History
- Horror
- Lifestyle

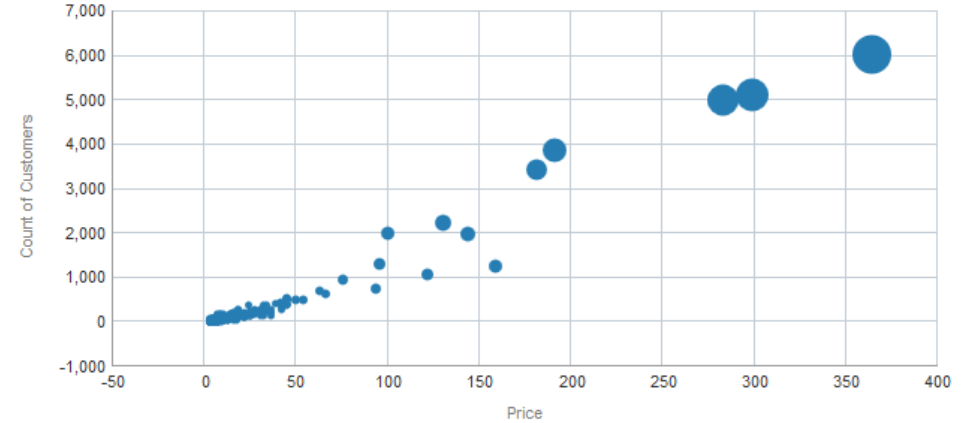
Customer Country

--Select Value--

Apply Reset

Movie Title	Price	Count of Customers	Revenue
Match Point	364	6,007	\$2,187,569
Risky Business	299	5,106	\$1,524,141
Cabin Fever	283	4,999	\$1,412,617
The Time Machine	191	3,846	\$734,740
Dirty Dancing	181	3,421	\$619,509
Sweet Home Alabama	129	2,255	\$291,684
Four Rooms	143	1,965	\$281,545
Memento	158	1,279	\$202,683
Candyman	100	2,005	\$199,498
Braveheart	122	1,065	\$129,493
Gladiator	96	1,303	\$124,671
American Beauty	75	973	\$72,732
Schindler's List	93	765	\$70,908
Casablanca	63	696	\$43,702

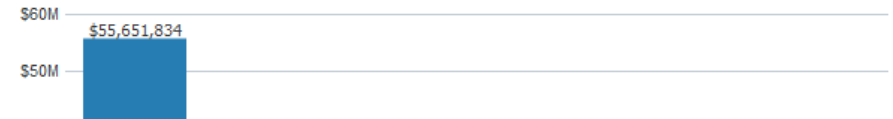
Revenue, Count of Customers, Price



Revenue from different Customer segments



Revenue



Page Information (click to collapse or expand)

Description : This demonstrates Impala connectivity to data residing in HDFS using Impala ODBC driver included in OBIEE installation (e.g. <BI Home>/bi/common/ODBC/Merant/7.1.4/lib/ARimpala27.so). The Impala SQL dialect supports a range of standard elements, plus some extensions for Big Data use cases related to data loading and data warehousing. Impala provides a high degree of compatibility with the Hive Query Language (HiveQL). Refer Impala SQL Language Reference documentation from Cloudera for more details.

Month

- (All Column Value)
- 2007-07
- 2007-08
- 2007-09
- 2007-10

Customer Country

--Select Value--

Income Level

--Select Value--

Genre Name

- NULL
- Action
- Adventure
- Animation
- Biography
- Comedy

Top Customers with Purchases

Purchase data from HDFS joined with Customer data in Oracle tables

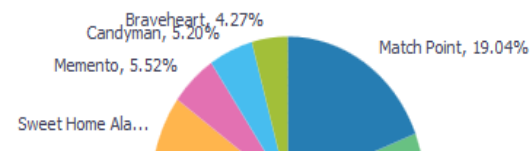
Customer Name	Customer Country	Income Level	Gender	Marital Status	Purchase Cnt
Wilfredo Wilder	United States	B: 30,000 - 49,999	Male	S	24
Carlo Gonzales	United States	B: 30,000 - 49,999	Male	S	23
Maryann Powers	United States	A: Below 30,000	Female	M	21
Phineas Benton	Hungary	C: 50,000 - 69,999	Male	M	21
Jita Nilini	India	A: Below 30,000	Male	S	19
Harold Abe	Japan	C: 50,000 - 69,999	Male	S	18
Heriberto Jong	United Kingdom	B: 30,000 - 49,999	Male	S	18
Jerrold Small	United States	A: Below 30,000	Male	S	18
Rafael Trujillo	United States	E: 90,000 - 109,999	Male	S	18
Tariyel De Vivero	Mexico	B: 30,000 - 49,999	Male	S	18



Top Movies Purchased

Generated by joining fact data from HDFS with movie attributes in Oracle relational tables

Movie Title	Release Year	No of Purchases	Revenue
Match Point	2005	183	364
Risky Business	1983	150	299
Cabin Fever	2002	142	283
The Time Machine	2002	104	207
Dirty Dancing	1987	91	181
Four Rooms	1995	81	161
Sweet Home Alabama	2002	66	131
Memento	2000	53	158
Candyman	1992	50	99
Braveheart	1995	41	122



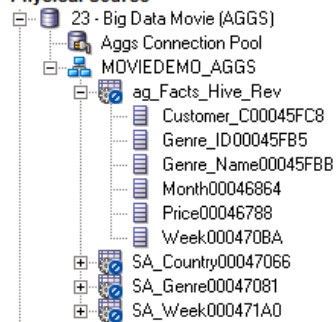
▲ Page Information (click to collapse or expand)

► **Description** : BI Server can act as a federation and aggregation engine that prepare and save results of advanced analytical from slower Hadoop tools to a high performance database such as Oracle DB. Below is an example of using aggregation persistence functionality of OBIEE to query hive tables and saving results into Oracle DB. The standard agg persistence functionality in OBIEE automatically increments its BI metadata and remaps columns so that subsequent queries (and its rollups) are returned from agg store directly and way more faster then Hive/Impala queries.

Sample aggregate creation script used for this example

```
create aggregates "ag_Facts_Hive_Revenue"
for
  "12 - Big Data Movie"."Facts (Hive)"("Price")
at levels (
  "12 - Big Data Movie"."H2 Customer"."Country",
  "12 - Big Data Movie"."H1 Movie"."Genre",
  "12 - Big Data Movie"."H0 Time"."Week")
using connection pool
  "23 - Big Data Movie (AGGS)".Aggs Connection Pool"
in
  "23 - Big Data Movie (AGGS)".."MOVIEDEMO_AGGS";
```

Physical Source



Customer Country --Select Value-- Select Metric Revenue (Aggregated) # of Activities (Not Aggregatec

Report sourcing from ORCL aggregate tables - Revenue Fact pre-aggregated from Hive

Genre Name	Revenue (Aggregated) - Year 2012														Total
	Week 27	Week 28	Week 29	Week 30	Week 31	Week 32	Week 33	Week 34	Week 35	Week 36	Week 37	Week 38	Week 39	Week 40	
Action	51	36	41	59	108	56	43	50	115	56	70	76	64	28	850
Adventure	27	19	33	22	86	39	40	36	67	29	30	22	27	15	490
Animation	8	12	3	2	15	9	8	18	24	16	13	2	9		138



▲ Page Information (click to collapse or expand)

► **Description** : This page highlights the difference in function shipping of analytical functions into physical SQL queries generated to Hive versus Oracle Big Data SQL. Query here is to find top 10 customers based on their number of activities on the Movie website. Logical Query to BI Sever is very similar in both cases, with Hive/Impala or Oracle Big Data SQL. However, physical queries against Hive doesn't include TOPN function, there by forcing the BI server to retrieve the granular data and then perform TOPN operation on its side. Whereas with Oracle Big Data SQL, the computation and filtering happens on the database side there by drastically improving the query performance.

Logical SQL query

```
SELECT
  0 s_0,
  "M - Big Data Movie"."Customer"."Cust ID" s_1,
  "M - Big Data Movie"."Facts (Hive)". "Count of Activities" s_2
FROM "M - Big Data Movie"
WHERE
(TOPN("Facts (Hive)". "Count of Activities",10) <= 10)
ORDER BY 1, 2 ASC NULLS LAST
FETCH FIRST 5000001 ROWS ONLY
```

BI Server Physical Query Log

```
[2015-01-14T01:51:37.572+00:00]----- Sending query to database named 20 - Big Data Movie (Hive)
(id: <<6354935>>), connection pool named Hive ODBC Connection Pool: []
select 0 as c1,
       D1.c2 as c2,
       D1.c1 as c3
from
  (select count(T290848.activity) as c1,
         T290848.custid as c2
   from
     movieapp_log_v T290848
   group by T290848.custid
  ) D1
]]
[2015-01-14T01:52:24.39+00:00]----- Query Status: Successful Completion
[2015-01-14T01:52:24.39+00:00]----- Rows 2260, bytes 54240 retrieved from database
[2015-01-14T01:52:24.39+00:00]----- Physical query response time 46.409 (seconds)
[2015-01-14T01:52:24.39+00:00]----- Physical Query Summary Stats: Number of physical queries 1,
Cumulative time 46.409, DB-connect time 0.000 (seconds)
[2015-01-14T01:52:24.39+00:00]----- Rows returned to Client 10
[2015-01-14T01:52:24.40+00:00]----- Logical Query Summary Stats: Elapsed time 46.478,
```

Logical SQL query

```
SELECT
  0 s_0,
  "M - Big Data Movie"."Customer"."Cust ID" s_1,
  "M - Big Data Movie"."Facts (BDSQL)". "Count of Activities" s_2
FROM "M - Big Data Movie"
WHERE
(TOPN("Facts (BDSQL)". "Count of Activities",10) <= 10)
ORDER BY 1, 2 ASC NULLS LAST
FETCH FIRST 5000001 ROWS ONLY
```

BI Server Physical Query Log

```
[2015-01-14T02:04:40.74+00:00]----- Sending query to database named 22 - Big Data Movie (BDSQL)
(id: <<6355850>>), connection pool named BDSQL Connection Pool: []
WITH
SAWITH0 AS (select count(T290878.ACTIVITYID) as c1,
              T290878.CUSTID as c2
 from
  MOVIEDEMO.MOVIEAPP_ACTIVITY_LOG T290878 /* MOVIEAPP_LOG_BDSQL_SRC */
 group by T290878.CUSTID),
SAWITH1 AS (select 0 as c1,
                 D1.c2 as c2,
                 D1.c1 as c3,
                 Rank() OVER ( ORDER BY D1.c1 DESC NULLS LAST ) as c4
 from
  SAWITH0 D1)
select D1.c1 as c1, D1.c2 as c2, D1.c3 as c3 from ( select D1.c1 as c1,
                 D1.c2 as c2,
                 D1.c3 as c3
 from
  SAWITH1 D1
 where ( D1.c4 <= 10 )
 order by c1, c2 ) D1 where rownum <= 5000001
```


Log Analysis | Log Details

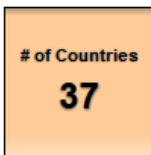
Page Information (click to collapse or expand)

Description : This report analyses log files generated on accessing an http server. Oracle Big Data SQL is used to read these log files residing in HDFS. The log files contain the host ipaddress from where the request is originating. This ip is joined to a lookup table in Oracle database to determine the country to which this ip belongs

Month 2014-Aug 2014-Jul 2014-Jun 2014-Oct 2014-Sep

Apply Reset

Log Analysis



2014-Aug

Country	# of Requests
United States	103,700
Canada	2,457
China	1,241
Japan	366
Netherlands	86
Thailand	76
Germany	23

of Requests

