# INTERSEC BENCHMARK

## High Performance for Fast Data & Real-Time Analytics
## Part I: Vs Hadoop

intersec

# BENCHMARK VS HADOOP

## (STAND ALONE OR COMBINED)

### Intersec solution in a Redhat Openstack NFV framework complements Hadoop platform and proves 100x more performant.

The combination of Intersec solution with Redhat Openstack NFV framework provides you with a flexible Big Data platform that lets you harness new and legacy data sources and drastically improve your responsiveness, whether you're using Hadoop applications or not. This document presents the tests led by Redhat and Intersec during the summer 2015, evaluating the performance of their combined solution in different environments, either in a stand-alone configuration or as a complement to an existing Hadoop infrastructure.

# I. INTERSEC INTEGRATION WITH REDHAT OPENSTACK NFV FRAMEWORK

Most service providers operate heterogenous networks with multiple, diverse data sources that prove challenging for legacy data management systems to stream and analyse at the required pace. The Intersec exclusive technology provides an efficient streaming analytics solution adapted to both real-time and historical data, whatever their types, volumes and sources. Redhat, as the expert of open technologies made safe and secure, combines the innovation of Openstack with the stability, security and support needed for production use. Together, Intersec and Redhat Openstack NFV framework deliver a high-performance solution compatible with Hadoop-based architectures.

This combination is ideal for high velocity businesses across all industries: telecommunications, financial services, public utilities, on-line retail or manufacturing, which sometimes have already developed some applications on Hadoop Framework and wish to keep them running. Intersec solution on Redhat framework enables them to boost their efficiency and to easily extend their scope to new use cases.

# II. SCOPE

Performance tests were run on both loading time and geomarketing use cases to measure Intersec and Hadoop technologies relative performances when run in a standalone mode or in combination with one another (reflecting the need of interworking with existing Hadoop applications).

### Loading time

A preliminary benchmark consisted in loading 1.42 TB of data (13 billions events) generated by customers' mobiles, on both Hadoop and Intersec frameworks, to compare the time required by both platforms. The sample of anonymized data extracted from a live environment as CSV files contained the following fields:

- Timestamp
- Subscriber identifier number (IMSI)
- Event type (call, SMS, location update, …)
- Cell identifier

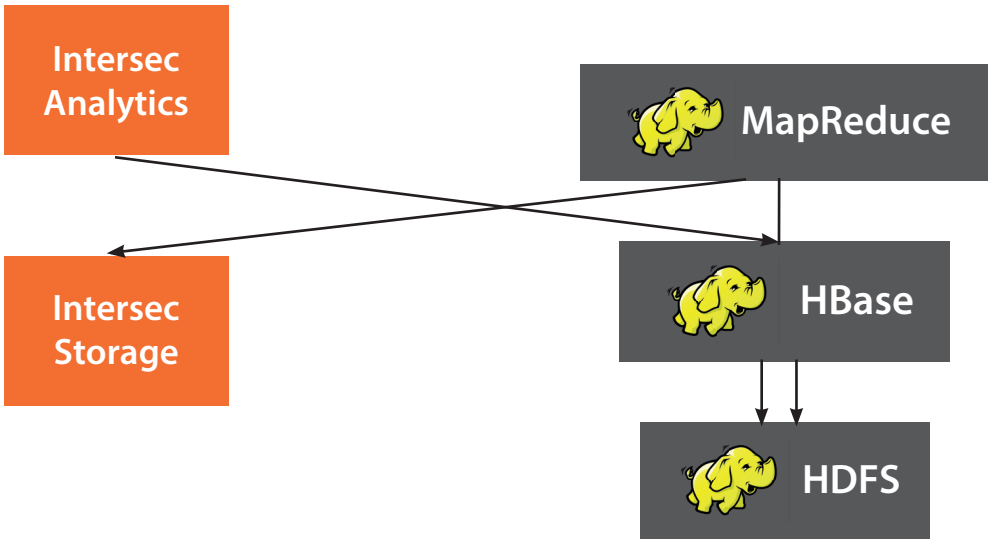
### Geomarketing use cases

This dataset enabled to run several geomarketing analysis on different combinations of Intersec and Hadoop technologies. One query consisted in simply counting the number of distinct customers inside a geographical zone and within a specific time-stamp range. Another one processed this result to compute individual trajectories and build so-called "origin-destination" (OD) matrices answering the question "how many customers went from (a list of) locations O to (a list of) locations D during this period of time?

Different measurements were made:
1. Time to count unique subscribers (several scales tested)
2. Time to count unique subscribers in a specific area (several scales and sizes of areas tested)
3. Time to process an OD matrix from end-to-end

## Technical environment

These tests were executed on different combinations between Intersec and Hadoop technologies to compare their performance:




- Intersec analytics solution running in standalone configuration
- Hadoop running in standalone configuration
- Hadoop running on top of Intersec, allowing to reuse Hadoop applications, in order to benefit from Intersec optimized storage technology.
- Intersec analytics running on top of Hadoop Hbase to show interoperability between Intersec technology and the existing Hadoop implementation.

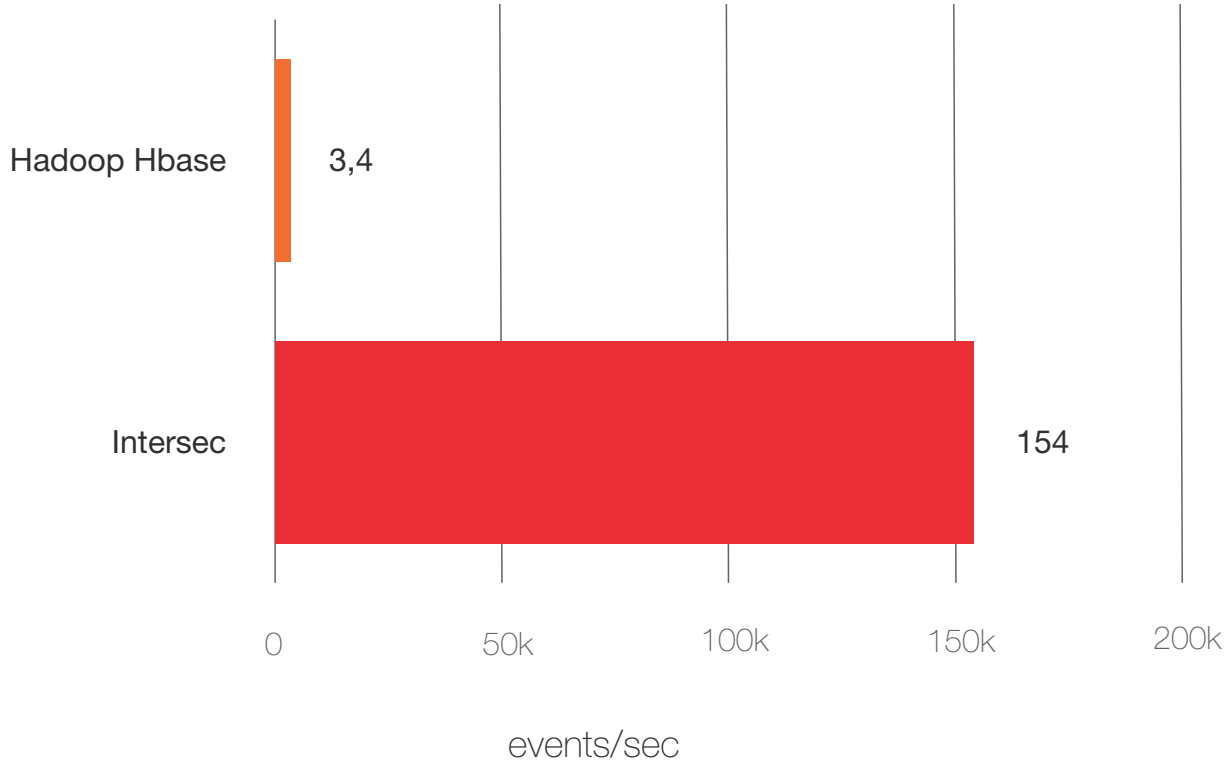
For each configuration, several queries were performed, from simple ones to a more complex sequence of queries leading to the computation of OD matrices. These queries are explained in more detail in section III.
All solutions were deployed on four virtual machines, each with the following configuration:
10 vCPUs
60GB of RAM
500GB of disk storage.
The storage was divided into four different hosts.

# III. RESULTS

The details regarding the tests and their results are in the appendix section.

## A. Loading 45 times faster with Intersec storage technology

As it is always possible to trade-off performance between loading and queries, it was important to measure the time to load and index the full dataset which would be used for further geomarketing queries.



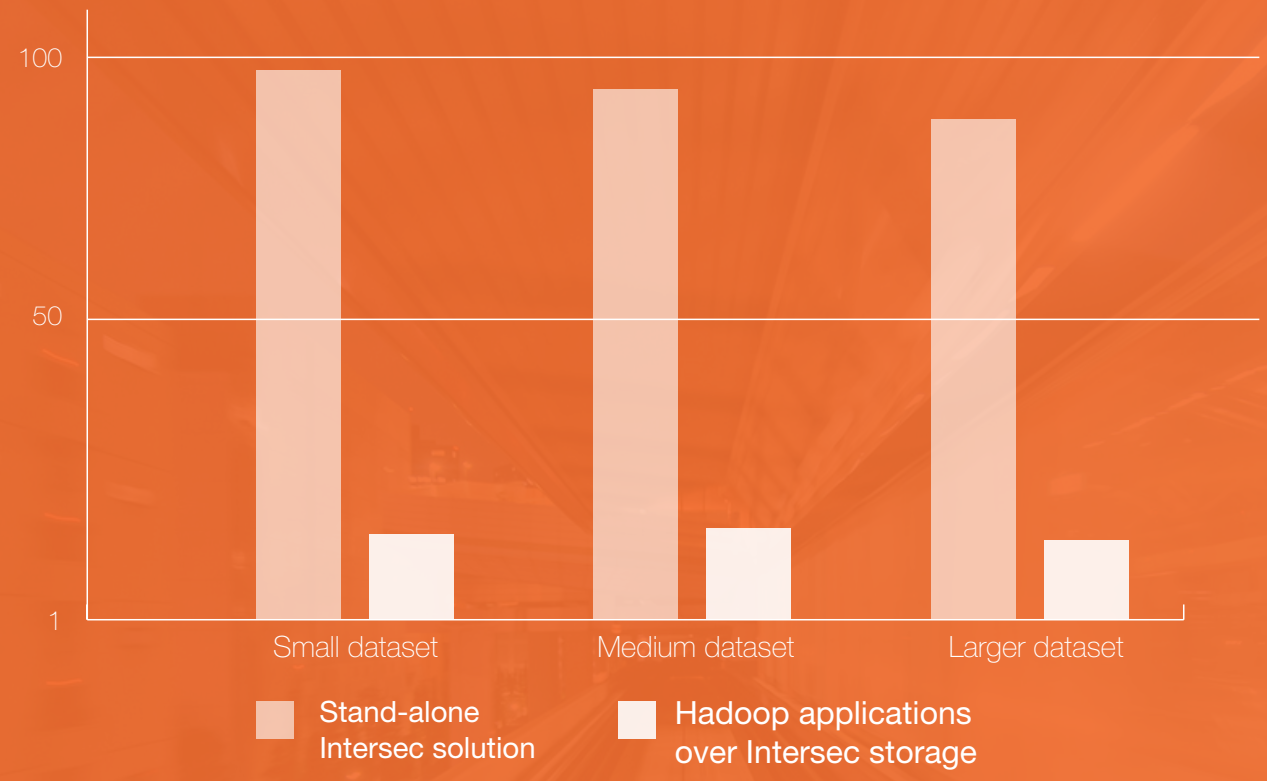

*Compared loading speed per host for Intersec and Hadoop Hbase technologies, measured on a 1,42TB dataset (larger is better)*

**45x** FASTER

2 and a half hours were necessary to complete this task using Intersec technology while it took over 106 hours on Hadoop Hbase.

## B. Intersec technology queries unique customers in a given time range more than 80 times faster

This query consisted in counting how many unique customers could be identified within different time ranges, represented by subdatasets of three different sample sizes: 118, 900 and 1300 million events, in order to measure the impact of the scale on performance.




*Compared querying speed to count the number of unique customers out of different sizes of dataset (stand-alone Hadoop configuration taken as a reference)*

Whatever the size of the dataset, Intersec standalone configuration is the quickest to execute the query, taking advantage of its efficient storage and index, as well as its dedicated querying process. It is on average over 80 times faster than a standalone Hadoop configuration.

When Hadoop applications are querying on top of Intersec storage instead of a classical Hbase framework, their performance is boosted by a factor of 13 or more.

The solution works but performances are not so interesting.

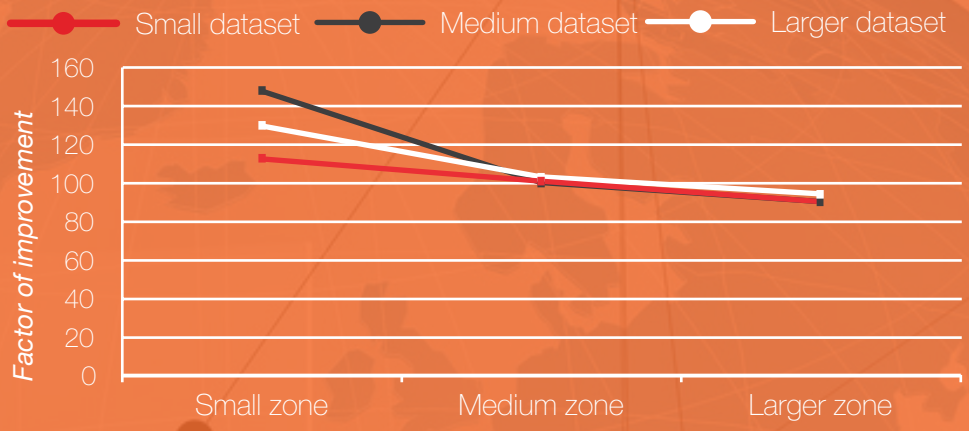
## C. Intersec technology improves query times by an average of 100 times to count unique customers in a given time range and zone

This query added a new dimension to the previous tests as it consisted in counting the number of unique customers in a given time range but also within a given geographical zone.
Technically this adds a difficulty to the previous test as the events to be selected are now scattered within the storage, while they were contiguous in the previous tests.
We chose to run the test over three sizes of zones (25 cells, 285 cells, and 4 872 cells), which combined with the three sizes of dataset gave nine possible combinations (see detailed results in the appendix) (cf. graph-1 p8):

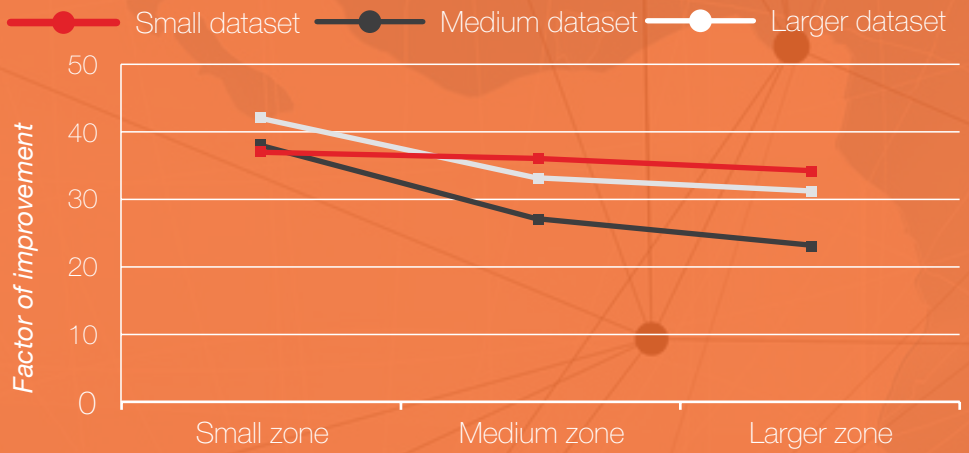
**100x** INTERSEC STANDALONE CONFIGURATION IS FASTER

In all cases Intersec standalone configuration is 100 times faster than its Hadoop equivalent.
We observed a great influence of data sparsity. The more scattered the queried events, the more significant the advantage of Intersec technology over Hadoop. On small zones with large dataset (the most difficult query to perform), Intersec solution performs nearly 150 times faster than Hadoop. This may be explained by the lack of efficient multi-criteria indexation on our Hbase configuration, which causes larger data scan on Hbase and longer query times.

Running Hadoop application over Intersec storage also showed a great factor of improvement - between 25 and 42, with larger improvement again in small zones (cf. graph-2 p8):




*Standalone Intersec technology querying speed to count the number of unique customers in a specific zone (different sizes of dataset and zones tested). Standalone Hadoop configuration taken as reference*

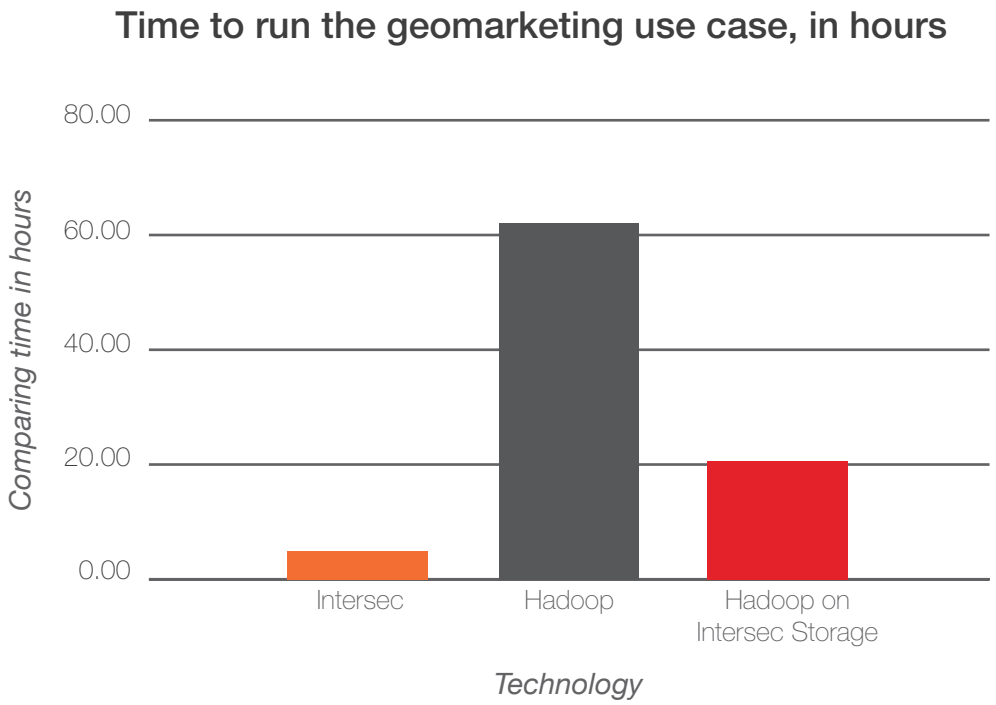



*Running Hadoop application over Intersec storage technology boosts performance by 30*

## D. Running an end-to-end geomarketing use case showed 14 times quicker on Intersec framework.

We chose to measure the time of execution of an end-to-end scenario, based on large zone and medium dataset, corresponding to 28 million initial events. The use case consisted in:
identifying the list of customers present in the zone (710 732 entries);
extracting the corresponding events and compute trajectories;
using the trajectories to build an origin/destination matrix on this timeframe.

Detailed results for each step are displayed in the appendix. Overall, the Intersec solution ran the whole scenario in less than 5 hours, while it took days to run it on a standalone Hadoop framework:




The results are similar to what was found before: the main part of the process relies on the ability to compute trajectories from events, reinforcing Intersec's comparative advantage.
We noticed a lower ratio in this test (around 14 here), as the extraction is based on customers rather than cells (the index is therefore heavier).

# IV CONCLUSION

These tests showed that, in different use cases of geomarketing, ranging from loading data, simple queries to a complete end-to-end scenario, Intersec technology implemented on Redhat Openstack NFV framework significantly improves processing times versus Hadoop technology by a factor of 100 in most queries.

This improved performance enables users of the platform to get quicker results and to replay a query several times, changing some elements to fine-tune their study.

The tests also showed perfect compatibility between Intersec technology and Hadoop. For companies already using Hadoop applications, adding Intersec storage technology can improve their efficiency by a factor from 10 to 40 times.

# APPENDIX

In the following tables, we will use H for Hadoop and I for Intersec.

## A. Count unique customers in a given time range

3 different queries, respectively based on 118, 900 and 1 305 millions of events.

| Timerange | Nb events | I/I | H/H | I/H | H/I |
|---|---|---|---|---|---|
| **T1 (small dataset)** | 117 535 190 | 170 | 15 043 | 28 474 | 1 083 |
| **T2 (medium dataset)** | 899 850 717 | 1 278 | 111 246 | 132 451 | 7 521 |
| **T3 (larger dataset)** | 1 304 799 454 | 3 170 | 259,940 | N/A | 19 044 |

*Querying times to retrieve the number of unique customers in a given time range, for the four configurations, in seconds*

## B. Count unique customers in a given time range and zone

As the size of the sample could have an impact on the querying speed, we created different zones:
Z1 (small zone): 25 cells,
Z2 (medium zone): 285 cells,
Z3 (larger zone): 4 872 cells.
The following table lists the number of events found in each {time range, zone} couple, and the resulting computing speeds.

| Time range, Zone | Nb events | I/I | H/H | I/H | H/I |
|---|---|---|---|---|---|
| **T1, Z1** | 49 384 | 4 | 451 | 812 | 12 |
| **T1, Z2** | 157 468 | 13 | 1315 | 1921 | 36 |
| **T1, Z3** | 3 498 244 | 55 | 4972 | 7812 | 144 |
| **T2, Z1** | 443 012 | 28 | 3625 | 5615 | 85 |
| **T2, Z2** | 2 628 587 | 106 | 10931 | 17545 | 323 |
| **T2, Z3** | 28 148 108 | 182 | 17182 | 21223 | 555 |
| **T3, Z1** | 644 346 | 77 | 11427 | 18471 | 301 |
| **T3, Z2** | 3 412 000 | 326 | 32611 | 51679 | 1201 |
| **T3, Z3** | 39 194 582 | 1135 | 101221 | NA | 4394 |

*Querying speeds to retrieve the number of unique customers in a given time range, for the four configurations, in seconds*

## C. Full geomarketing use case

We applied the different steps of the full use case, based on the T2, Z3 couple, corresponding to 28,1 millions of initial events.
The following table reports the computing times for each of the steps, as well as the resulting global times to perform the use case.

| Step | I/I | H/H | I/H | H/I |
|---|---|---|---|---|
| **Identify Unique Users** | 201 | 17 156 | 21 312 | 563 |
| **Extract events, compute trajectories** | 14 544 | 202 146 | NA | 74 221 |
| **Generate OD matrix** | 275 | 301 | NA | 299 |
| **Total** | 15 020 | 219 603 | NA | 75 083 |

*Computing times to perform the full geomarketing use case, for the four configurations, in seconds. (smaller is better)*