TECHNOLOGY DETAIL

RED HAT DATA ANALYTICS INFRASTRUCTURE SOLUTION

TABLE OF CONTENTS

Give data scientists and data analytics teams access to their own clusters without the unnecessary cost and complexity of duplicating Hadoop Distributed File System (HDFS) datasets.

Rapidly deploy and decommission analytics clusters on demand with Red Hat private cloud infrastructure.

Share analytics datasets in Red Hat Ceph Storage object stores, avoiding unnecessary duplication of large datasets and eliminating delays from data hydration and destaging.

Use space-efficient erasure coding for data protection, saving up to 50% of per-cluster storage costs over 3x replication.[1]

facebook.com/redhatinc
@redhat
linkedin.com/company/red-hat

redhat.com

[1] *Testing by Red Hat and QCT, 2017-2018, redhatstorage.redhat.com/2018/07/02/why-spark-on-ceph-part-3-of-3/*

INTRODUCTION

Traditional data analytics infrastructure is under stress due to the enormous volume of captured data and the need to share finite resources among teams of data scientists and data analysts. New technology and computing capabilities have created a revolution in the amounts of data that can be retained and in the kinds of insights that can be garnered. However, divergent objectives have emerged between teams who want their own dedicated clusters, and the underlying data platform teams who would prefer shared datacenter infrastructure. In response, some data platform teams are offering teams their own Apache Hadoop Distributed File System (HDFS) clusters. Unfortunately, this approach results in expensive duplication of large datasets to each individual cluster, as HDFS is not traditionally shared between different clusters.

Public cloud providers like Amazon Web Services (AWS) offer a more compelling model, where analytics clusters can be rapidly deployed and decommissioned on demand—with each having the ability to share datasets in a common object storage repository. Emulating this architectural pattern, several leading large enterprises have used Red Hat® private cloud platforms to deploy agile analytics clusters, sharing datasets in a common AWS Simple Storage Service (S3)-compatible Ceph® object storage repository.

Based on the pioneering success of these efforts, Red Hat, Quanta Cloud Technology (QCT), and Intel Corporation sought to quantify the performance and cost ramifications of decoupling compute from storage for big data analytics infrastructure. They were aided by the S3-compatible Hadoop S3A filesystem client connector that can be used with an S3-compatible object store to augment or replace HDFS. Using a shared data lake concept based on Red Hat Ceph Storage, compute workloads and storage can be independently managed and scaled, providing multitenant workload isolation with a shared data context.

RED HAT DATA ANALYTICS INFRASTRUCTURE SOLUTION

After conversations with more than 30 companies,[2] Red Hat identified a host of issues with sharing large analytics clusters. Teams are frequently frustrated because someone else's job prevents their job from finishing on time, potentially impacting service-level agreements (SLAs). Moreover, some teams want the stability of older analytics tool versions on their clusters, whereas their peers might need to load the latest and greatest tool releases. As a result, many teams demand their own separate and specifically tailored analytics cluster so that their jobs are not competing for resources with other teams.

With traditional Hadoop, each separate analytics cluster typically has its own dedicated HDFS datastore. To provide access to the same data for different Hadoop/HDFS clusters, the platform team frequently must copy very large datasets between the clusters, trying to keep them consistent and up-to-date. As a result, companies maintain many separate, fixed analytics clusters (more than 50 in one company Red Hat interviewed[2]), each with its own redundant data copy in HDFS containing potentially petabytes of data. Keeping datasets updated between clusters requires an error-prone maze of scripts. The cost of maintaining 5, 10, or 20 copies of multipetabyte datasets on the various clusters is prohibitive to many companies in terms of capital expenses (CapEx) and operating expenses (OpEx).

[2] *Red Hat conversations with early adopters, 2017-2018.*

Organizations that have outgrown a single analytics cluster have several choices:

- Get a bigger cluster for everyone to share.
- Give each team its own dedicated cluster, each with a copy of potentially petabytes of data.
- Give teams the ability to spin up and spin down clusters that can share datasets.

The Red Hat analytics infrastructure solution offers a novel approach, yielding the ability to rapidly spin-up and spin-down clusters while giving them access to shared data repositories. Now, technologies like Red Hat OpenStack® Platform, Red Hat OpenShift® Container Platform, and Red Hat Ceph Storage can be combined with industry-standard servers to bring many of these same benefits to on-premise analytics infrastructure.

THE EVOLUTION OF ANALYTICS INFRASTRUCTURE

With the growing prevalence of cloud-based analytics, data scientists and analysts have grown accustomed to dynamically deploying clusters on services like AWS. By design, these clusters have access to shared datasets, without time-consuming HDFS data hydration periods after initializing a new cluster or destaging cycles upon cluster termination. Specifically, AWS allows users to rapidly launch many analytics clusters on Amazon Elastic Compute Cloud (Amazon EC2) instances, and share data between them on Amazon Simple Storage Service (Amazon S3).

Many analysts now expect these same capabilities on-premise, and we are witnessing an evolution in the ways that on-premise analytics infrastructure is contemplated and deployed.

- Historically, most organizations initially ran analytics on bare-metal systems, hoping to get the most performance from the systems while taking advantage of data locality from HDFS.
- Many organizations are now running analytics in virtual machines, provisioned by OpenStack, allowing for some of the flexibility of cloud-based analytics infrastructure.
- Some organizations are employing containers for on-premise analytics deployments, increasing both performance and flexibility.

Analytics testing with Red Hat Ceph Storage and the Hadoop Distributed File System (HDFS) was performed during 2017 and 2018 by Red Hat and QCT using QuantaPlex T21P-4U servers at a QCT lab facility. The same systems and testing environment were used for both Ceph and HDFS testing.[3]

Both of the latter deployment methods typically call upon Ceph Storage as a software-defined object store. This functionality is enabled by the Hadoop S3A filesystem client connector, used by Hadoop to read and write data from Amazon S3 or a compatible service. With the S3A filesystem client connector, Apache Spark and Hadoop jobs and queries can run directly against data held within a shared S3-compatible datastore.

BENEFITS OF A SHARED DATA REPOSITORY ON RED HAT CEPH STORAGE

Red Hat Ceph Storage is a natural choice for organizations that want to provide an S3-compatible shared data repository experience to their analysts on-premise. Based on testing by Red Hat and QCT,[3] supporting Spark or Hadoop analytics on Ceph provides a number of benefits over traditional HDFS.

- **Reduced CapEx by reducing duplication.** Petabytes of redundant storage capacity are often purchased to store duplicate datasets in HDFS. With Ceph-based data storage, this redundancy can be reduced or eliminated while allowing access to the same datasets by multiple clusters.
- **Reduced CapEx by improving data durability efficiency.** HDFS typically requires 3x replication for data protection. In addition to replication, Ceph also supports erasure coding, potentially reducing the CapEx of purchased storage capacity by up to 50% due to improved data durability efficiency.
- **Right-sized CapEx infrastructure costs.** Traditional HDFS clusters can suffer from over-provisioning of either compute or storage resources. Shared Ceph data storage promotes right-sizing of compute needs (in terms of virtual central processing unit [vCPU] or random access memory [RAM]) independently from storage needs (in terms of throughput or capacity).

[3] *Testing by Red Hat and QCT, 2017-2018, redhatstorage.redhat.com/2018/07/02/why-spark-on-ceph-part-3-of-3/*

The Hadoop Distributed File System (HDFS) was designed for performance and data locality through a tightly-coupled relationship between compute and storage within the cluster. This tight coupling can be a challenge when different groups seek to share a Hadoop cluster, resulting in large-scale data movement, proliferation of separate clusters, and escalating costs. Moreover, organizations that want to add storage also have to add more compute, causing one or the other to be under or over utilized. Data locality too can suffer in rapidly growing Hadoop clusters, potentially lowering performance as compute resources have to travel further to find the data they need.

- **Reduced OpEx and risk.** Scripting and scheduling dataset copies between HDFS instances to maintain consistency and allow multiple analytics clusters access to the same data can be costly. With Ceph, clusters can retain access to the same data without these costs, while reducing the risk of human error.
- **Accelerated insights from new data science clusters.** When launching new clusters, having to copy or rehydrate data into a new cluster can significantly delay important analysis. Analyzing data in place within a shared Ceph data repository can dramatically reduce time to insight.
- **Support for different tool and version needs of diverse data teams.** Different teams have different needs regarding tool and distribution versions. With a shared Ceph datastore, users of each cluster can choose the Spark or Hadoop toolsets and versions appropriate to their jobs, without disrupting users from other teams requiring different tools and version.

SOLUTION COMPONENTS

To document the ability of Ceph to serve as a shared datastore, Red Hat, QCT, and Intel designed a reference architecture that builds on successful early adopter deployments and now forms the basis for the Red Hat data analytics infrastructure solution.

- **Red Hat OpenStack Platform** is a cloud computing platform that virtualizes resources from industry-standard hardware, organizes those resources into clouds, and manages them. In the context of this solution, it provides on-demand provisioning of virtualized analytics clusters.
- **Red Stack OpenShift Container Platform** is a reliable, enterprise-grade platform that combines the industry-leading container orchestration engine with advanced application build and delivery automation features. Red Hat OpenShift Container Platform is an optional element of this solution for those who are interested in containerizing Spark clusters.
- **Red Hat Ceph Storage** is an open, massively scalable storage solution for modern workloads like cloud infrastructure and data analytics. It provided the shared object repository for all testing described in this document.

TESTING ENVIRONMENT OVERVIEW

In addition to simplistic tests such as TestDFSIO, Red Hat wanted to run analytics jobs that were representative of real-world workloads. The TPC Benchmark DS (TPC-DS) was used for ingest, transformation, and query jobs. TPC-DS generates synthetic datasets and provides a set of sample queries intended to model the analytics environment of a large retail company with sales operations from stores, catalogs, and the web. Its schema has tens of tables, with billions of records in some tables. The benchmark defines 99 preconfigured queries, from which the Red Hat team selected the 54 most input/output (I/O)-intensive queries for testing. Categories of tests included:

- **Bulk ingest.** Bulk load jobs, simulating high-volume streaming ingest at 1PB+/day.
- **Ingest.** Composed of MapReduce jobs.
- **Transformation.** Apache Hive or Spark structured query language (SQL) jobs that can convert plain text data into Apache Parquet or optimized row columnar (ORC) compressed formats.
- **Query.** Hive or Spark SQL jobs frequently run in batch or noninteractive mode, as these tools automatically restart jobs.

- **Interactive query.** Apache Impala or Presto jobs that offer interactive SQL analytic query capabilities.
- **Merge or join.** Hive or Spark SQL jobs joining semistructured clickstream data with structured web sales data.

QCT servers powered by Intel® Xeon® processors were used for testing both the analytics and storage clusters. The QCT QuantaPlex T21P-4U server was used in Red Hat testing as it provides large storage capacity in a small space. It also includes Peripheral Component Interconnect Express (PCIe) Generation 3 (Gen3) slots, allowing Non-volatile Memory Express (NVMe) solid-state drives (SSDs). In Red Hat testing, the Intel Solid State Drives Data Center Family was used for Ceph journaling to accelerate both I/O per second (IOPS) and throughput performance. Capable of delivering up to 780TB of storage in just one 4-rack-unit (4U) system, the QuantaPlex T21P-4U efficiently serves the most demanding cloud storage environments. As shown in Figure 1, a Spark/Hadoop compute cluster was connected with an object store based on Red Hat Ceph Storage, communicating via the Amazon S3A filesystem client connector. The clusters were configured as follows:

- **Analytics cluster.** 20 Spark/Hadoop compute hosts with one or more YARN or Impala clusters (480 physical compute cores/960 logical compute cores).
- **Storage cluster.** 20 Ceph storage hosts configured into a single, shared Ceph object store. Collectively, the storage servers featured 700 hard disk drives (HDDs) and 40 NVMe SSDs.

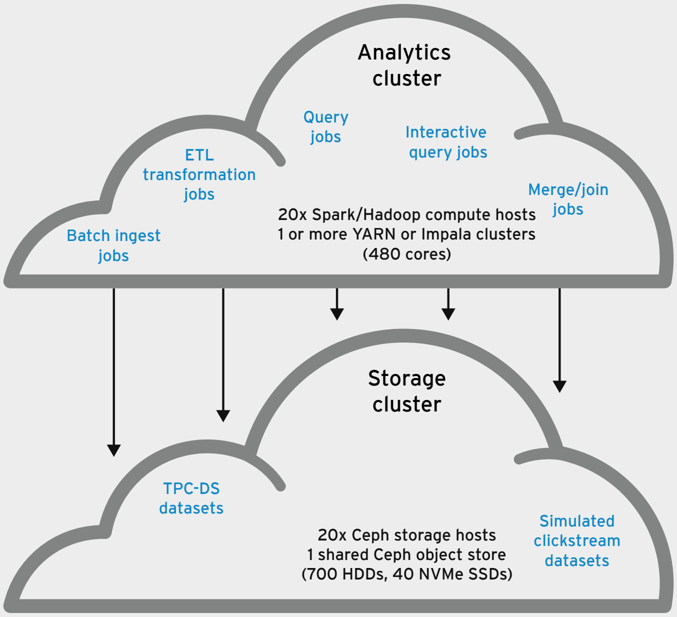


Figure 1. For testing, Red Hat utilized 20 QCT servers configured as Spark/Hadoop nodes running against a shared Ceph object store running Red Hat Ceph Storage on 20 QCT servers.

RELATIVE COST AND PERFORMANCE COMPARISON

Many factors contribute to overall solution cost. Storage capacity is frequently a major component of the price of a big data solution, so it was a simple proxy in the price/performance comparison in this study. The sections that follow provide a summary and details of the results achieved.

FINDINGS SUMMARY

The primary factor affecting storage capacity price in Red Hat's comparison was the scheme used for data durability (data protection). With 3x data replication, an organization needs to buy 3PB of raw storage capacity to yield 1PB of usable capacity. In contrast, erasure coding (EC) 4:2 data durability requires only 1.5PB of raw storage capacity to obtain the same 1PB of usable capacity. As of this writing, the primary data durability scheme used by HDFS is 3x replication. Support for HDFS erasure coding is emerging but is still experimental in several distributions.

In contrast, Ceph has supported both erasure coding and 3x replication as data durability schemes for years. All of the early adopters that Red Hat worked with are using erasure coding for cost-efficiency reasons. As such, most of the Red Hat testing was conducted with erasure-coded clusters (EC 4:2). Some tests were run with Ceph 3x replication used in the storage cluster to provide an equivalent comparison to HDFS for those tests.

Using storage capacity as a proxy for cost as described above, Figure 2 provides a price/performance summary for eight different workloads described in the sections below (keyed to the illustration). Each of the eight workloads was run with both HDFS and Ceph storage. As expected, the cost of the storage capacity was either the same (when Ceph was configured for 3x replication), or 50% less (when Ceph was configured for erasure coding). For example, workload No. 8 exhibited similar performance with either Ceph or HDFS, but the Ceph storage capacity price was 50% of the HDFS storage capacity price. In the cases of workload No. 1 and No. 2, Ceph 3x replication was used and the relative cost was equivalent to that of HDFS. In the other examples, erasure coding was used, resulting in a 50% cost savings for storage capacity with somewhat lower performance.

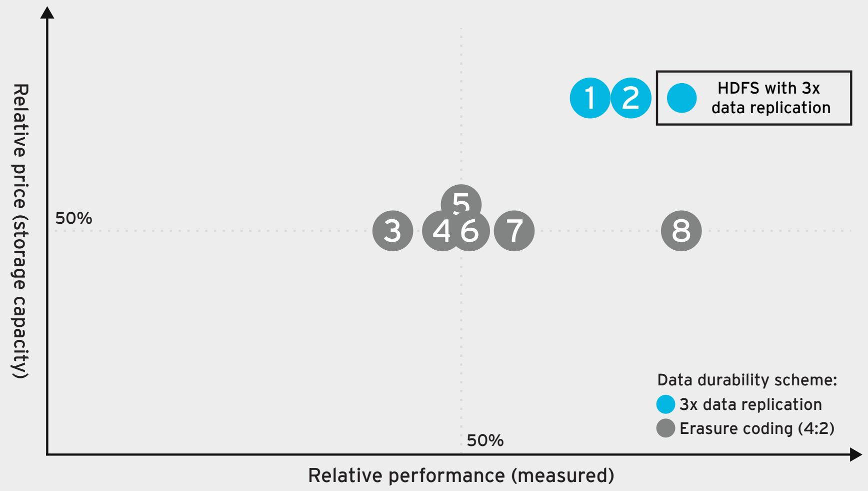


Figure 2. Relative price/performance comparison, based on results from eight different workloads

WORKLOAD DETAILS

Workloads tested are represented in Figure 2 and included the following:

1. **TestDFSIO—3x replication.** This workload was a simple test to compare aggregate read throughput via TestDFSIO. It performed comparably between HDFS and Ceph when Ceph also used 3x replication. When Ceph used erasure coding (EC 4:2), the workload performed better than it did for either HDFS or Ceph with 3x replication for lower numbers of concurrent clients (fewer than 300). With more client concurrency, however, the workload performance on Ceph EC 4:2 dropped due to spindle contention.[4]

2. **SparkSQL single-user (1x)—3x replication.** This workload compared the Spark SQL query performance of a single user executing a series of queries (the 54 TPC-DS queries described above). Engineers found that the aggregate query time was comparable when running against either HDFS or Ceph with 3x replicated storage but that it doubled with Ceph EC 4:2 storage.

3. **Impala multiuser (10x)—EC 4:2.** This workload compared Impala query performance of 10 users, each executing a series of queries concurrently. The 54 TPC-DS queries were executed by each user in a random order. As illustrated, the aggregate execution time of this workload on Ceph with EC 4:2 storage was 57% slower than on HDFS. However, price/performance was nearly comparable, as the HDFS storage capacity costs were twice those of the Ceph configuration.

4. **Mixed workload (Spark, Impala)—EC 4:2.** This mixed workload featured concurrent execution of a single user running Spark SQL queries (54), 10 users each running Impala queries (54 each), and a dataset merge/join job enriching TPC-DS web sales data with synthetic clickstream logs. As illustrated in Figure 2, the aggregate execution time of this mixed workload on Ceph with EC 4:2 storage was 48% slower than on HDFS. However, price/performance was nearly comparable, as the HDFS storage capacity costs were twice those of Ceph with EC 4:2 storage.

5. **TestDFSIO (100% writes)—EC 4:2.** This workload was a simple test to compare aggregate write throughput via TestDFSIO. It performed 50% slower on average on Ceph EC 4:2 storage than on HDFS, across a range of concurrent clients/writers. However, price/performance was nearly comparable, as the HDFS storage capacity costs were twice those of Ceph with EC 4:2 storage.

6. **SparkSQL single-user (1x)—EC 4:2.** This workload compared Spark SQL query performance of a single user executing the 54 TPC-DS queries—identical to test No. 2, but using erasure-coded storage. In testing, the aggregate query time was comparable when running against either HDFS or Ceph with 3x replicated storage. The aggregate query time doubled when running against Ceph with EC 4:2 storage. Again, price/performance was nearly comparable when running against Ceph with EC 4:2 storage, due to high HDFS storage capacity costs.

7. **SparkSQL data migration—EC 4:2.** This workload featured the enrichment (merge/join) of TPC-DS web sales data with synthetic clickstream logs, subsequently writing the updated web sales data. This workload was 37% slower on Ceph with EC 4:2 storage than on HDFS. However, price/performance was favorable for Ceph, as the HDFS storage capacity costs were twice those of Ceph with EC 4:2 storage.

[4] *A single read with erasure coded 4:2 storage requires four disk accesses vs. a single disk access with 3x replicated storage, resulting in more contention with a greater number of concurrent clients.*

8. **SparkSQL multiuser—EC 4:2.** This workload compared the Spark SQL query performance of 10 users, each executing a series of queries concurrently. (The 54 TPC-DS queries were executed by each user in a random order.) As illustrated in Figure 2, the aggregate execution time of this workload on Ceph with EC 4:2 storage was roughly comparable to that of HDFS, despite requiring only half of the storage capacity costs. Price/performance for this workload thus favors Ceph by a factor of 2. For more insight into this workload performance, see Figure 3. In this box-and-whisker plot, each dot reflects a single Spark SQL query execution time. As each of the 10 users concurrently executes 54 queries, there are 540 dots per series. The three series shown are Ceph with EC 4:2 storage (green), Ceph with 3x replication (red), and HDFS with 3x replication (blue). The Ceph with EC 4:2 replication box shows median execution times comparable to those of HDFS, and shows more consistent query times in the middle two quartiles.

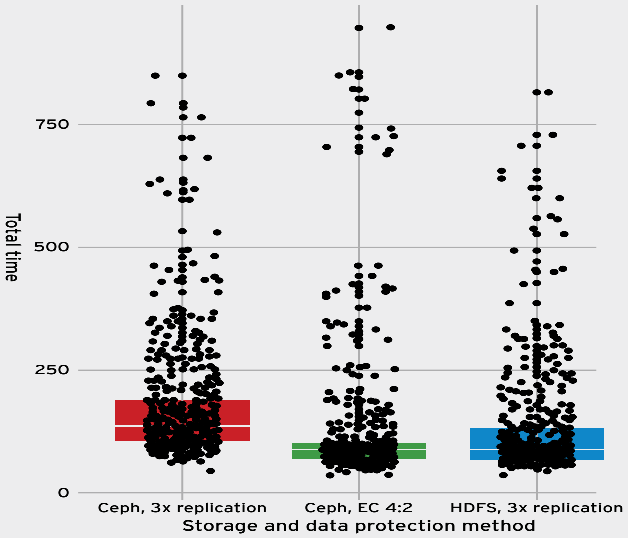


Figure 3. Multiuser Spark query set results

24-HOUR INGEST

In addition to the testing described above, Red Hat conducted testing over a 24-hour time period to illustrate Ceph cluster sustained ingest rate. For these tests, Red Hat used a variation of the lab as described above. Testing measured a raw ingest rate of approximately 1.3 pebibytes (PiB) per day into a Ceph cluster with EC 4:2 storage configured with 700 HDD data drives (Ceph Object Storage Daemons, or OSDs).

CONCLUSION

The Red Hat data analytics infrastructure solution, based on Red Hat Ceph Storage as a shared data repository for big data, addresses the growing pains of traditional data analytics infrastructure. It offers significant economic and operational benefits—and some performance trade-offs. Sharing storage between Apache Hadoop clusters clearly makes sense, and CapEx and OpEx can be reduced dramatically by eliminating petabytes of duplicated storage. At the same time, the approach gives different teams the flexibility to choose the exact analytics tools and versions they need. In a head-to-head comparison, Red Hat Ceph Storage can provide up to 50% storage infrastructure cost savings over HDFS through erasure coding, while allowing organizations to scale compute and storage infrastructure independently as needs dictate.

FOR MORE INFORMATION

Deploy flexible, on-premise analytics infrastructure using the Red Hat data analytics infrastructure solution by maximizing the power of Red Hat OpenStack Platform, Red Hat OpenShift Container Platform, and Red Hat Ceph Storage.

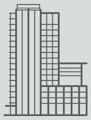
- For more information on the testing detailed in this piece, visit the three-part blog series Why Spark on Ceph?
- For a more detailed technical treatment of the subject of analytics infrastructure on Ceph, visit the 10-part technical blog series that begins with What about locality?
- To schedule a complimentary proof-of-concept (PoC) scoping workshop, led by experts at Red Hat Consulting, or to deploy a full PoC designed to address your individual analytics needs, visit Red Hat Consulting.

ABOUT QCT

Quanta Cloud Technology (QCT) is a global datacenter solution provider, combining the efficiency of hyperscale hardware with infrastructure software from a diversity of industry leaders to solve next-generation datacenter design and operation challenges. QCT serves cloud service providers, telecoms, and enterprises running public, hybrid, and private clouds. Product lines include hyperconverged and software-defined datacenter solutions as well as servers, storage, switches, and integrated racks with a diverse ecosystem of hardware component and software partners. QCT designs, manufactures, integrates, and services cutting-edge offerings via its own global network. The parent of QCT is Quanta Computer, Inc., a Fortune Global 500 corporation. www.QCT.io.

ABOUT RED HAT

Red Hat is the world's leading provider of open source software solutions, using a community-powered approach to provide reliable and high-performing cloud, Linux, middleware, storage, and virtualization technologies. Red Hat also offers award-winning support, training, and consulting services. As a connective hub in a global network of enterprises, partners, and open source communities, Red Hat helps create relevant, innovative technologies that liberate resources for growth and prepare customers for the future of IT.



facebook.com/redhatinc
@redhat
linkedin.com/company/red-hat

redhat.com
#F14280_1118

NORTH AMERICA	EUROPE, MIDDLE EAST, AND AFRICA	ASIA PACIFIC	LATIN AMERICA
1 888 REDHAT1	00800 7334 2835	+65 6490 4200	+54 11 4329 7300
	europe@redhat.com	apac@redhat.com	info-latam@redhat.com