

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



NGUYỄN HỮU THẢO THUẬN

**CÔNG NGHỆ BIG DATA
VÀ ỨNG DỤNG PHÂN TÍCH SỐ LIỆU KINH DOANH
CỦA TẬP ĐOÀN VIETTEL**

CHUYÊN NGÀNH : KHOA HỌC MÁY TÍNH

MÃ SỐ: 60.48.01.04

LUẬN VĂN THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. NGUYỄN ĐÌNH HÓA

HÀ NỘI - 2016

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Nguyễn Đình Hóa

Phản biện 1:

Phản biện 2:

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại
Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ... giờ ngày tháng năm

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Ngày nay, sự phát triển của Internet đã làm thay đổi mạnh mẽ cách thức hoạt động của các tổ chức. Các ứng dụng Web 2.0, mạng xã hội, điện toán đám mây đã một phần mang lại cho các tổ chức phương thức kinh doanh mới. Số lượng người sử dụng máy tính và các tài nguyên trực tuyến để xử lý công việc, giải trí, ... ngày càng tăng nhanh. Đặc biệt dữ liệu được tạo ra và truyền tải trên internet là vô cùng lớn, cụ thể: dữ liệu trên thế giới tăng gấp đôi sau mỗi 2 năm. Google có hơn 3 triệu máy chủ để xử lý hơn 1,7 nghìn tỷ lượt tìm kiếm trong một năm, các trung tâm dữ liệu tiêu thụ gần 1,5% điện năng trên toàn thế giới, có 571 website mới được tạo ra sau mỗi phút, dự đoán sẽ có 1/3 lượng dữ liệu trên thế giới sẽ được lưu trữ và truyền tải thông qua "đám mây" vào năm 2020, Twitter xử lý 7 TB dữ liệu mỗi ngày, Facebook xử lý 10 TB dữ liệu mỗi ngày, có 750 triệu bức ảnh được đăng tải lên Facebook mỗi 2 ngày, có hơn 247 tỷ email được gửi đi mỗi ngày, gần 80% email là thư rác, số lượng tin nhắn văn bản được gửi và nhận mỗi ngày vượt qua số lượng con người trên hành tinh này, 48 giờ video được đăng tải lên YouTube mỗi phút, tương đương lượng nội dung số dài 8 năm mỗi ngày[1].

Trong kỷ nguyên của IoT, các cảm biến được nhúng vào trong các thiết bị di động như điện thoại di động, ô tô, và máy móc công nghiệp,... việc chuyển dữ liệu định kỳ hoặc liên tục từ chiếc xe bạn lái về máy chủ tại chính hãng không còn là chuyện viễn tưởng nữa. Vấn đề chỉ còn là xử lý: kích thước, tốc độ, phương thức xử lý và kết quả đầu ra. Ford, GE hay Rolls Royce cùng rất nhiều hãng xe hơi khác đang đầu tư vào IoT. Điều tương tự cũng xuất hiện ở nhiều ngành khác, vốn là kịch bản tất yếu của khái niệm vạn vật kết nối. Hệ quả tất yếu là khối lượng dữ liệu số đang phình to ra với tốc độ chóng mặt. Khối lượng dữ liệu mới được tạo ra nhiều và nhanh đến mức mà hai năm gần đây nhất chiếm đến 90% khối lượng dữ liệu trên thế giới hiện nay. Những dữ liệu này tới từ mọi nơi. Ví dụ như từ những chiếc cảm biến để thu thập thông tin thời tiết, những thông tin được cập nhật trên các trang web mạng xã hội, những bức ảnh và video kỹ thuật số được đưa lên mạng, dữ liệu giao dịch của các hoạt động mua sắm trên mạng... dưới mọi hình thức khác nhau (có cấu trúc, phi cấu trúc, bán cấu trúc).

Theo một báo cáo của IDC, năm 2011, lượng dữ liệu được tạo ra trên thế giới là 1.8 ZB (ngàn tỷ byte), tăng gần 9 lần chỉ trong 5 năm. Năm 2012 là 2.8 ZB. Dự báo đến năm 2020 là 40 ZB. Dưới sự bùng nổ này, thuật ngữ Big Data được sử dụng để chỉ những bộ dữ liệu khổng lồ, chủ yếu không có cấu trúc, được thu thập từ nhiều nguồn khác nhau. Với

những tác động trong việc khám phá giá trị tiềm ẩn to lớn, Big Data đang được xem là một yếu tố mới quan trọng mang lại lợi ích cho các tổ chức trong nhiều lĩnh vực khác nhau. Các chuyên gia tài chính đánh giá đầu tư vào Big Data sẽ là yếu tố then chốt để đạt được lợi thế cạnh tranh. Chính vì những lợi ích to lớn mà Big Data có thể mang lại, nhiều tổ chức đã đầu tư mạnh vào việc nghiên cứu và ứng dụng vào xử lý khai thác Big Data [1].

Tại Tập đoàn Viễn thông Quân đội Viettel, cùng với việc mở rộng mạng lưới kinh doanh dịch vụ viễn thông toàn cầu, khối lượng dữ liệu tăng trưởng rất mạnh. Đặc biệt là số liệu kinh doanh: hóa đơn điện tử, giao dịch đầu nối, dữ liệu cước, ... Việc đầu tư vào nghiên cứu ứng dụng công nghệ Big Data để đưa ra các quyết định kinh doanh kịp thời và chính xác là rất cần thiết.

Từ nhu cầu thực tế đó, tác giả quyết định chọn đề tài **“Công nghệ Big Data và ứng dụng phân tích số liệu kinh doanh của Tập đoàn Viettel”** cho luận văn tốt nghiệp với mục đích nghiên cứu công nghệ Big Data và giải quyết bài toán xử lý số liệu kinh doanh tại Viettel.

Cấu trúc luận văn

Nội dung của luận văn được trình bày trong ba phần chính như sau:

1. Phần mở đầu
2. Phần nội dung: bao gồm ba chương

Chương 1: Làm rõ định nghĩa Big Data và hiện trạng ứng dụng khai thác xử lý Big Data ở Việt Nam và trên thế giới. Giới thiệu tổng quan về 3 giải pháp Big Data. Đề xuất sử dụng công nghệ Apache Hadoop để xây dựng module xử lý số liệu kinh doanh của Viettel.

Chương 2: Trình bày chi tiết công nghệ Hadoop.

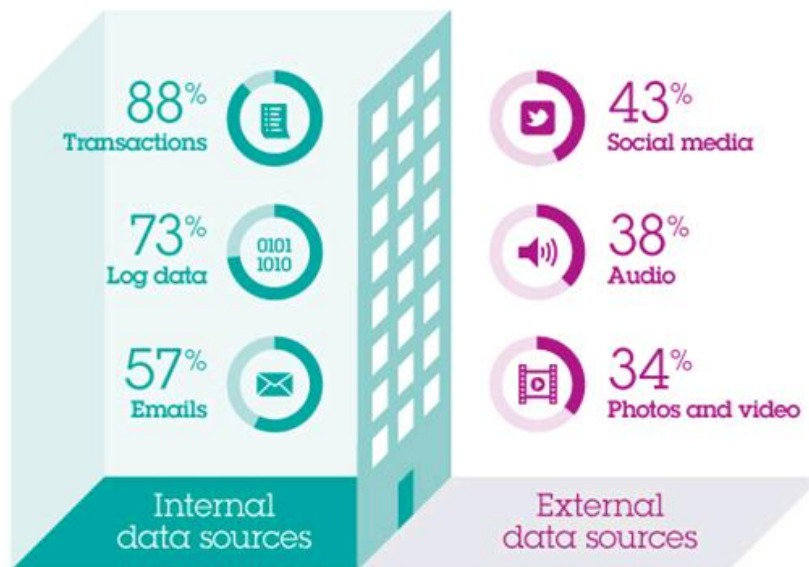
Chương 3: Trình bày xây dựng ứng dụng xử lý số liệu kinh doanh tại Viettel.

3. Phần kết luận

CHƯƠNG 1. TỔNG QUAN VỀ BIG DATA

1.1 Khái niệm về Big data

Big Data là một thuật ngữ dùng để mô tả các bộ dữ liệu có kích thước rất lớn, khả năng phát triển nhanh, rất khó thu thập, lưu trữ, quản lý và phân tích với các công cụ thống kê hay ứng dụng cơ sở dữ liệu truyền thống [2].



Hình 1.1 Thống kê các nguồn dữ liệu hiện nay[2]

1.2 Các đặc tính của việc xử lý Big Data

Thứ nhất là độ lớn dữ liệu (volume), nghĩa là dữ liệu sinh ra tự động có số lượng nhiều hơn rất nhiều so với dữ liệu truyền thống.

Thứ hai là tốc độ xử lý dữ liệu (Velocity), tức là dữ liệu lớn không đồng nghĩa với xử lý chậm.

Thứ ba là tính đa dạng dữ liệu (variety), tức là với việc thu thập từ nhiều nguồn dữ liệu khác nhau (web, mobile...)

Thứ tư là giá trị (value), đây là đặc trưng quan trọng nhất của Big Data, đề cập đến quá trình trích xuất các giá trị to lớn đang tiềm ẩn trong các bộ dữ liệu khổng lồ.

1.2.1 Ứng dụng Big Data trong tài chính ngân hàng, bảo hiểm

1.2.1.1 Quản lý rủi ro

1.2.1.2 Tư vấn Big Data và các ứng dụng liên quan

1.2.1.3 Các kỹ thuật thống kê trên dữ liệu lịch sử

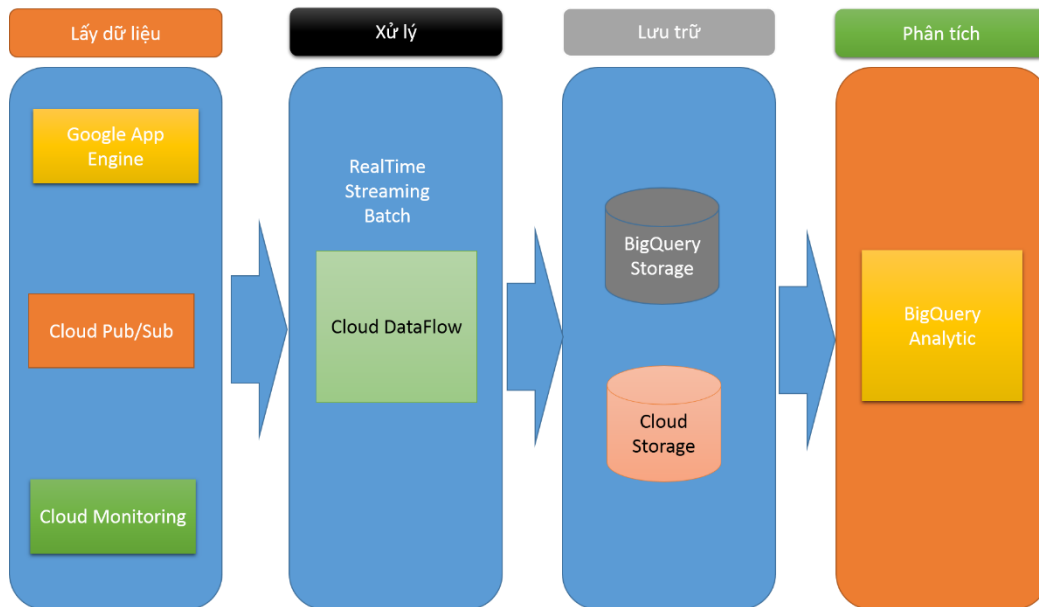
1.2.2 Thương mại

1.3 Hiện trạng khai thác Big Data trên thế giới và ở Việt Nam

1.4 Tổng quan về các giải pháp Big Data

1.4.1 Google Cloud Platform

1.4.1.1 Tổng quan



Hình 1.2 Mô hình kiến trúc mẫu hệ thống Big Data của google [3]

1.4.1.2 Các thành phần

1.4.1.2.1 Google App Engine

1.4.1.2.2 Google Cloud Pub/Sub

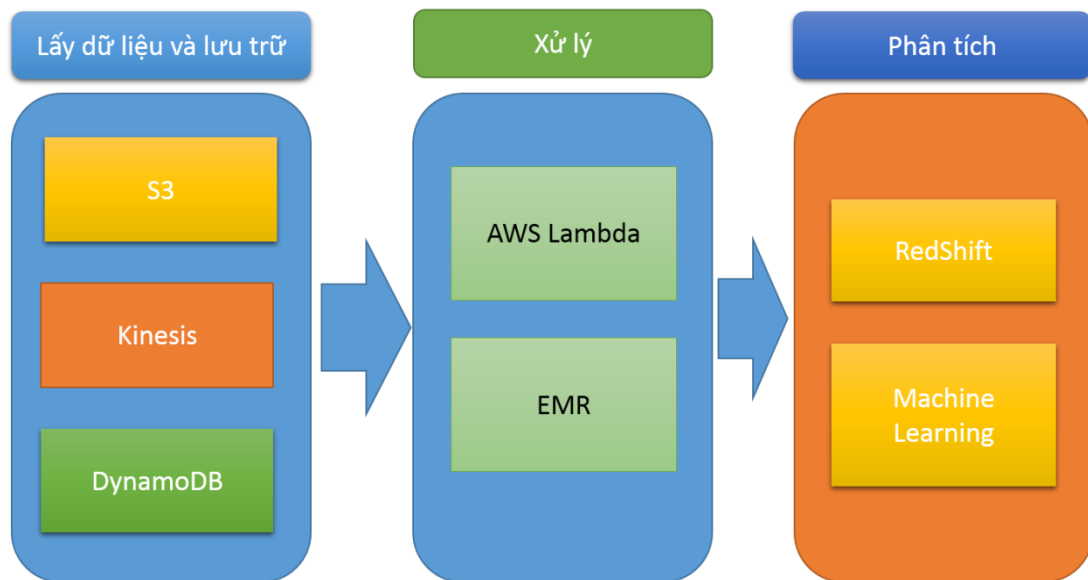
1.4.1.2.3 Google Cloud Monitoring

1.4.1.2.4 Google Cloud Storage

1.4.1.2.5 Google Cloud Dataflow

1.4.2 Amazon EMR

1.4.2.1 Giới thiệu tổng quan



Hình 1.3 Mô hình kiến trúc tích hợp Amazon webservice điển hình [3]

1.4.2.2 Các thành phần

1.4.2.2.1 Dịch vụ lưu trữ đơn giản của Amazon (S3)

1.4.2.2.2 Amazon Kinesis Streams

1.4.2.2.3 Amazon DynamoDB

1.4.2.2.4 AWS Lambda

1.4.2.2.5 Amazon EMR

1.4.2.2.6 Amazon Machine Learning

1.4.2.2.7 Amazon Redshift

1.4.3 Apache Hadoop

1.5 Kết luận chương

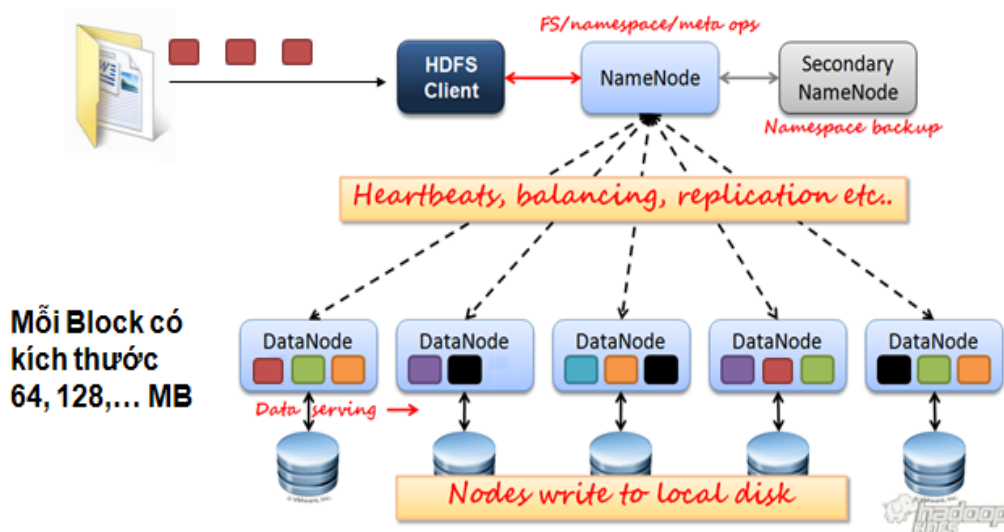
CHƯƠNG 2. CÔNG NGHỆ APACHE HADOOP

2.1 Giới thiệu về Hadoop

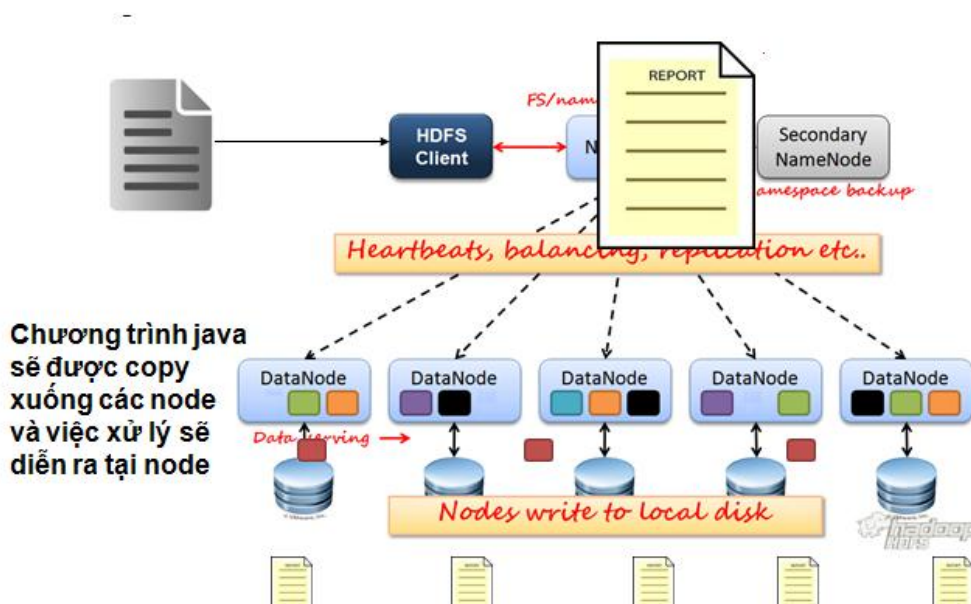
Hadoop có 2 thành phần chủ yếu là HDFS (Hadoop Distributed File System) và MapReduce [4].

Apache Hadoop định nghĩa: Apache Hadoop là một framework dùng để chạy những ứng dụng trên 1 cluster lớn được xây dựng trên những phần cứng thông thường. Hadoop hiện thực mô hình Map/Reduce

Wikipedia định nghĩa: Hadoop là một framework nguồn mở viết bằng Java cho phép phát triển các ứng dụng phân tán có cường độ dữ liệu lớn một cách miễn phí.



Hình 2.1 Mô hình Hadoop lưu trữ dữ liệu phân tán trên hệ thống Hadoop Distributed File System (HDFS)[4]



Hình 2.2 Mô hình Hadoop xử lý dữ liệu song song và phân tán trên các nút [4]

2.2 Các trình nền của Hadoop

2.2.1 NameNode

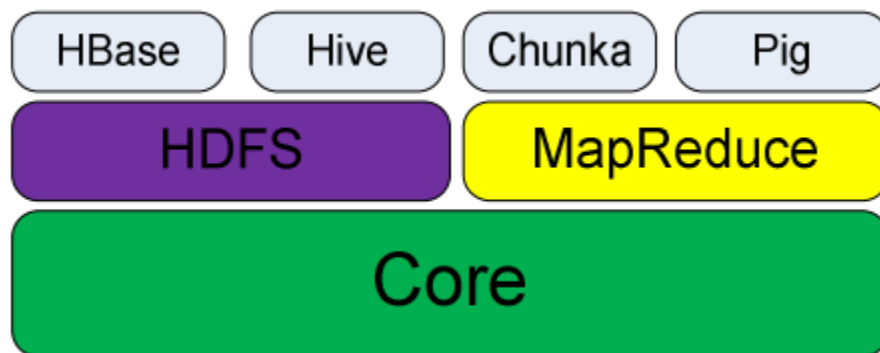
2.2.2 DataNode

2.2.3 Secondary NameNode

2.2.4 JobTracker

2.2.5 TaskTracker

2.3 Kiến trúc tổng thể Hadoop

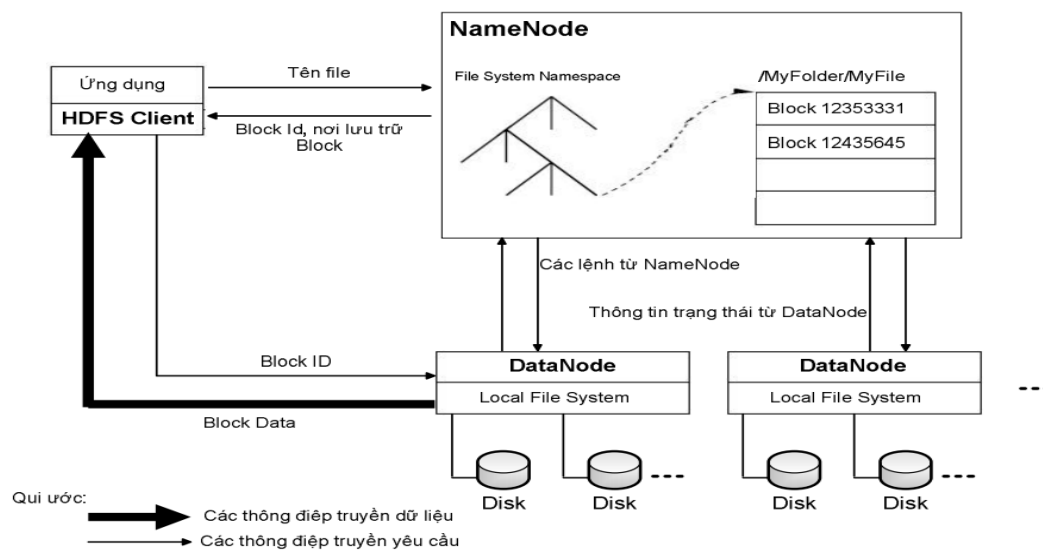


Hình 2.3 Mô hình kiến trúc tổng quát của Hadoop [4]

2.3.1 Hệ thống tập tin phân tán Hadoop (HDFS)

2.3.1.1 Kiến trúc các thành phần

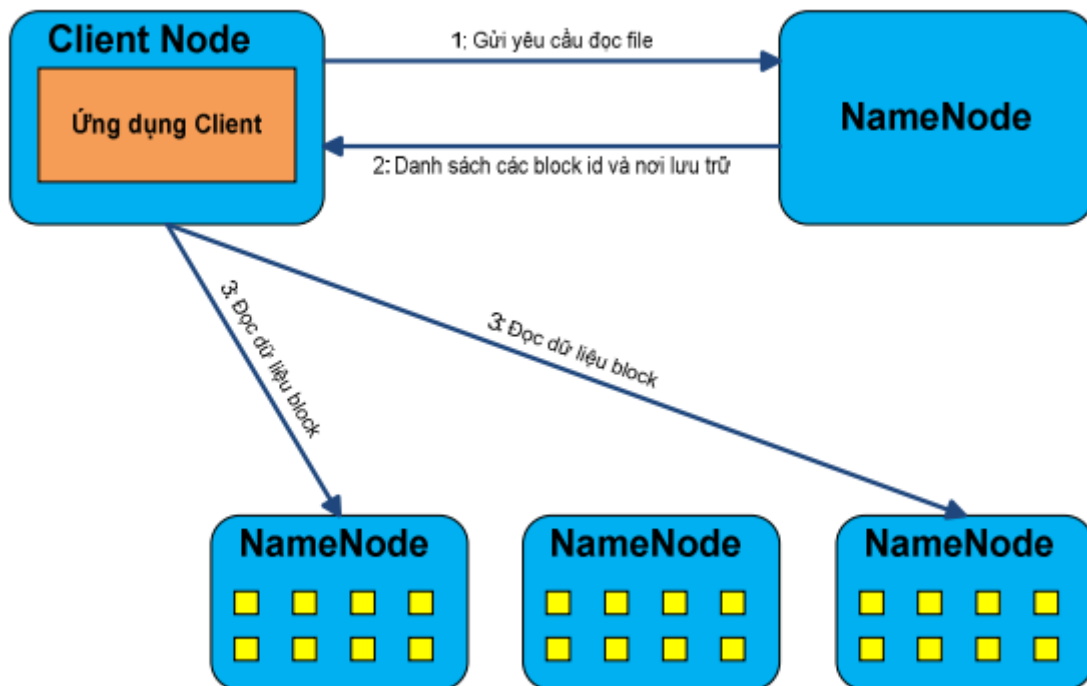
Kiến trúc của HDFS được thể hiện qua sơ đồ dưới đây:



Hình 2.4 Sơ đồ kiến trúc hệ thống HDFS [4]

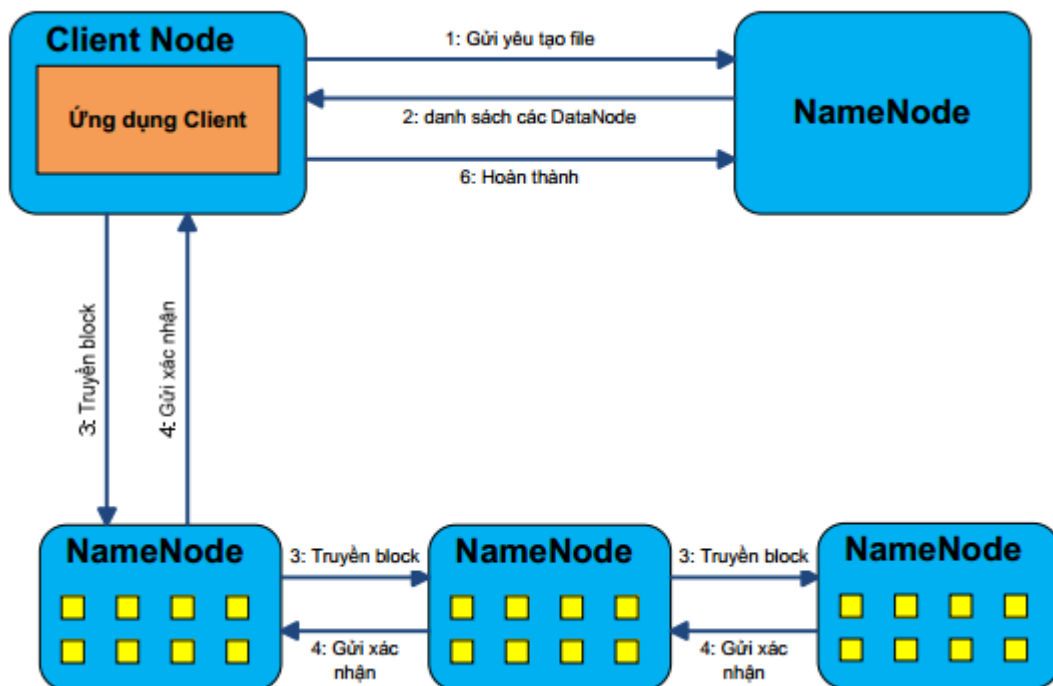
2.3.1.2 Cơ chế hoạt động

2.3.1.3 Quá trình đọc file



Hình 2.5 Sơ đồ quá trình client đọc một file trên HDFS [4]

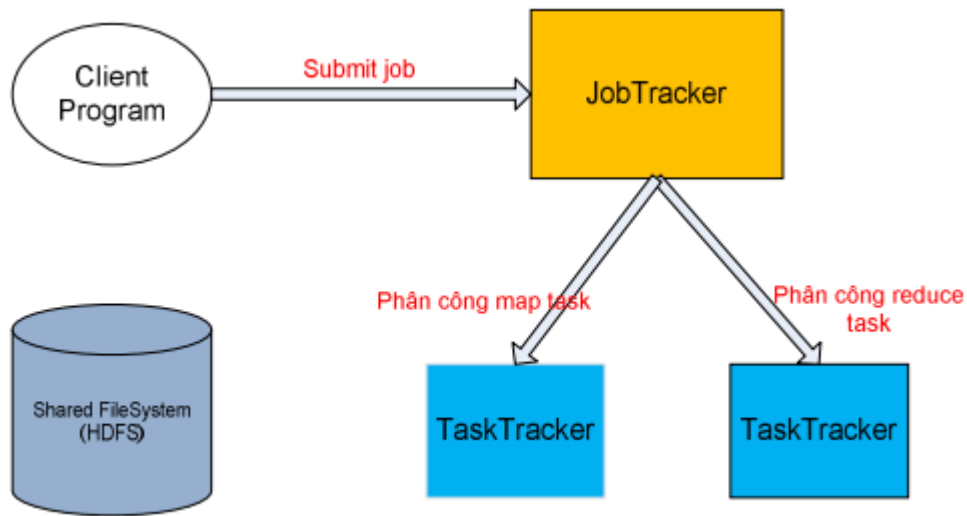
2.3.1.4 Quá trình ghi file



Hình 2.6 Sơ đồ quá trình ghi file trên HDFS [4]

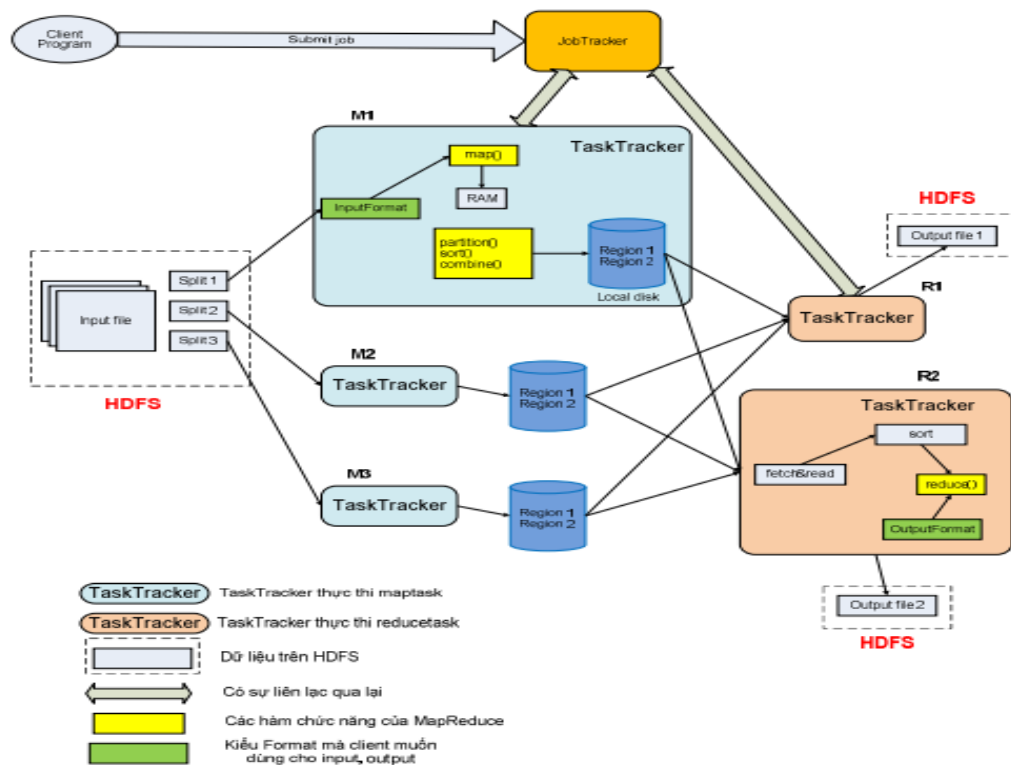
2.3.2 Hadoop MapReduce

2.3.2.1 Kiến trúc các thành phần



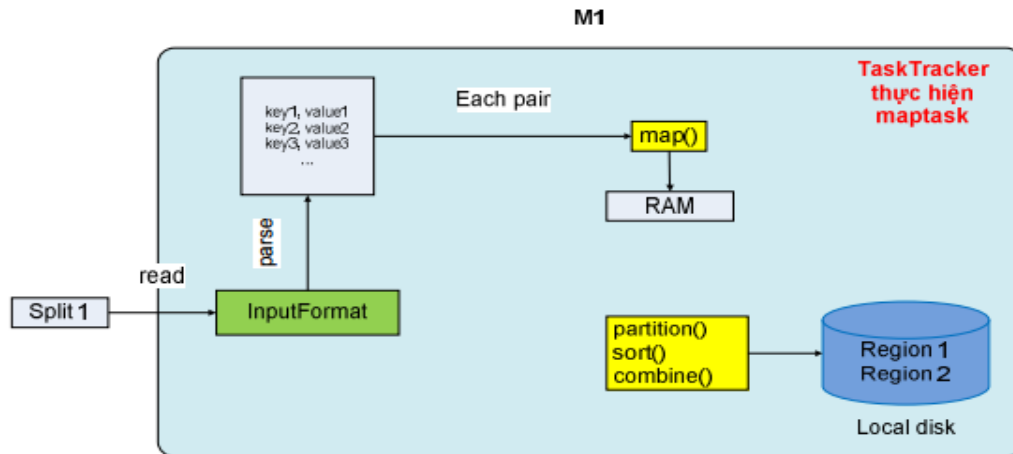
Hình 2.7 Sơ đồ thành phần Map Reduce [4]

2.3.2.2 Cơ chế hoạt động



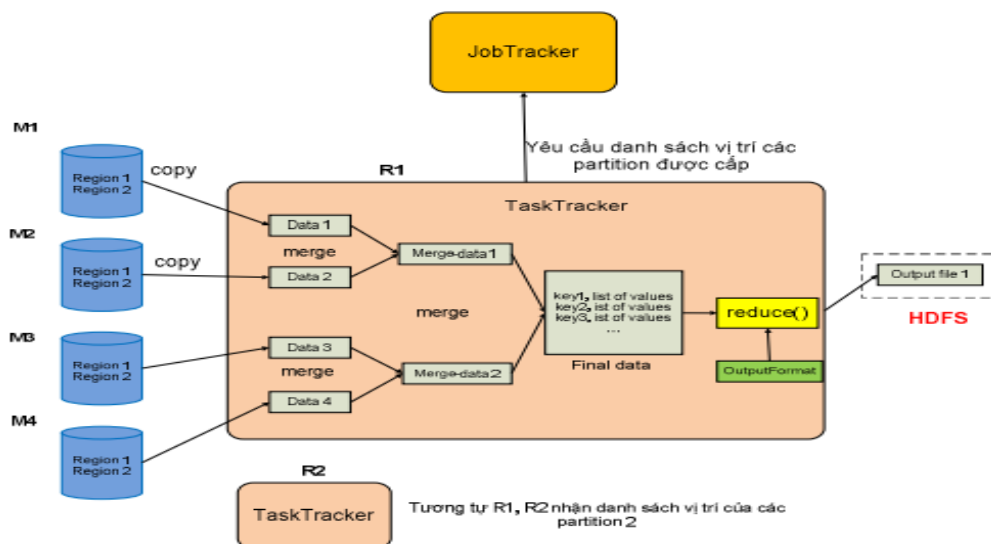
Hình 2.8 Sơ đồ luồng hoạt động Map Reduce [4]

Cơ chế hoạt động của maptask



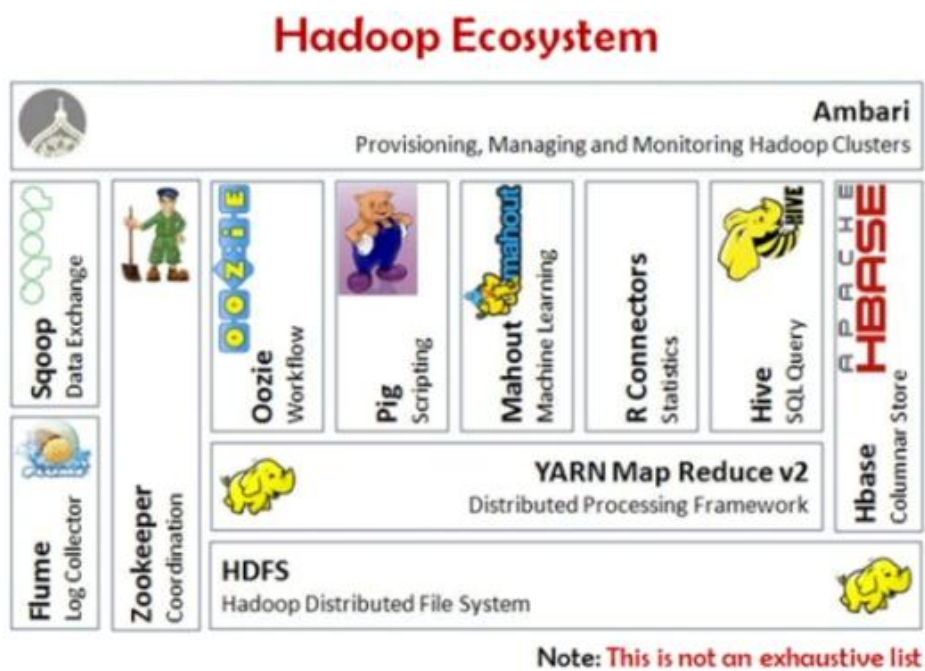
Hình 2.9 Sơ đồ luồng hoạt động của Map [4]

Cơ chế hoạt động Reduce Task



Hình 2.10 Sơ đồ luồng hoạt động của Reduce [4]

2.4 Hệ sinh thái các sản phẩm đi kèm Hadoop



Hình 2.11 Hình vẽ hệ sinh thái các sản phẩm trên Hadoop [4]

2.5 Kết luận chương

CHƯƠNG 3. ỨNG DỤNG

3.1 Đặt vấn đề bài toán ứng dụng

Về mặt sản phẩm:

- Xây dựng hệ thống Viettel Real-Time Big Data Analytics Platform triển khai cho thị trường Viettel Telecom (VTT) các thị trường Viettel đầu tư; đồng thời triển khai cho các doanh nghiệp, chính phủ bên ngoài.
- Hệ thống phát triển theo dạng thức là 1 nền tảng (platform) tổ chức, xử lý và khai thác dữ liệu. Cho phép triển khai linh hoạt và nhanh chóng các mô hình phân tích và kịch bản kinh doanh.
- Sản phẩm phải được kiểm chứng về mặt chức năng và hiệu năng ít nhất trên 1 thị trường mà Viettel đầu tư, là sở cứ để nghiệm thu sản phẩm.
- Làm chủ về mặt công nghệ để có thể dễ dàng thay đổi tính năng theo nhu cầu thị trường.

Về mặt kỹ thuật:

- Hệ thống xử lý phân tán (Cluster computing, Distributed File System), có khả năng mở rộng hệ thống theo chiều ngang khi lượng dữ liệu cần xử lý tăng lên.
- Hệ thống có khả năng xử lý thời gian thực (Stream processing) thấp nhất từ mức giây (seconds) trở xuống.
- Hệ thống hợp nhất nền tảng xử lý dữ liệu thời gian thực (Stream processing) và xử lý dữ liệu theo lô (Batch processing – như hiện tại của hệ thống ZTE BI và Viettel BI) dưới cùng một nền tảng công nghệ (Technology stack) In-memory Map Reduce/Cluster computing/Distributed File System.
- Hệ thống có khả năng chịu lỗi (Fault-tolerance). Khi một số phần tử (node) trong cụm (cluster) bị đổ vỡ (failed), hệ thống vẫn hoạt động bình thường.
- Hệ thống không tồn tại bất cứ điểm chết nào (single point of failure). Các cấu phần trong hệ thống được triển khai hoặc theo mô hình dự phòng (failover, active – passive) hoặc theo mô hình chia tải (load balancing, active – active).
- Hệ thống xây dựng trên nền tảng mã nguồn mở Apache Hadoop

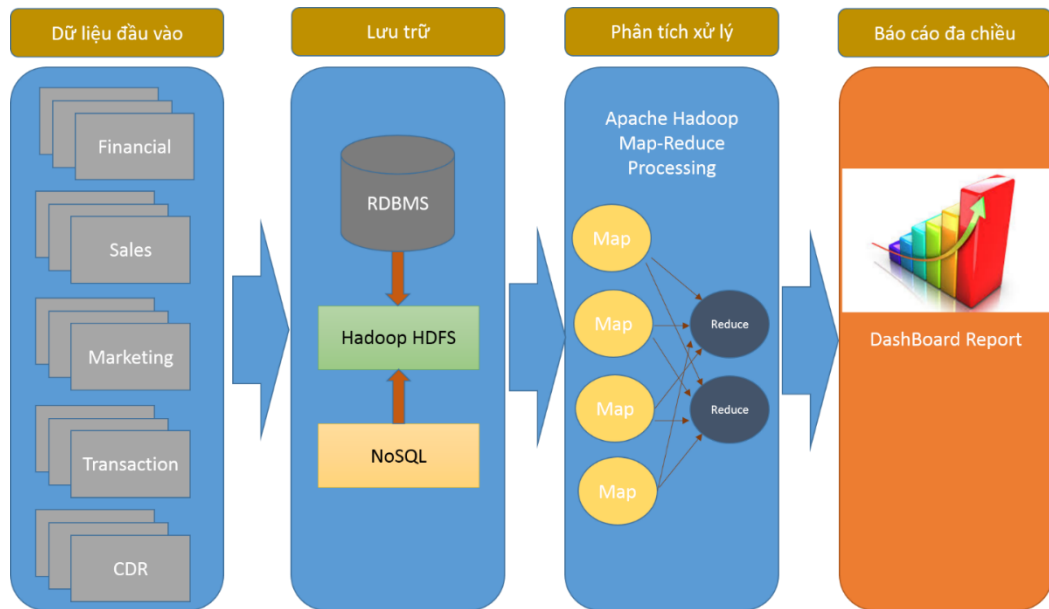
Về mặt kinh tế:

- Giảm chi phí đầu tư phần cứng, bản quyền phần mềm, chi phí triển khai khi mua sản phẩm của đối tác.

- Chủ động trong việc triển khai sản phẩm cũng như tùy biến theo thị trường và tích hợp với các sản phẩm khác của Viettel.

3.2 Xây dựng hệ thống

3.2.1 Mô hình kiến trúc tổng thể



Hình 3.1 Mô hình kiến trúc tổng thể hệ thống xử lý số liệu kinh doanh [5]

3.3 Mô tả dữ liệu đầu vào

Ví dụ về một mẫu file dữ liệu:

SERVICE_PK|F_VALUE|F_VALUE_MONTH|PRD_ID|DEP_ID|UNIT_ID

```
240|17|109|20140504|VTC|10
149|682897|2713950|20140504|VTC|12
150|4379|18038|20140504|VTC|12
242|7160|-1468|20140504|VTC|10
210|22842|98336|20140504|VTC|10
323|1199724|4472485|20140504|VTC|12
215|22116|-60112|20140504|VTC|10
230|23|156|20140504|VTC|10
28|674853|2686636|20140504|VTC|12
211|1817203|1817203|20140504|VTC|10
212|501242|-79130|20140504|VTC|10
231|4681|4681|20140504|VTC|10
```


232|1438|-403|20140504|VTC|10
 241|104790|104790|20140504|VTC|10
 319|3872008|3872008|20140504|VTC|12
 214|257204|257204|20140504|VTC|10
 151|14365|57541|20140504|VTC|12
 320|9848748|9848748|20140504|VTC|12

Mỗi file CDR chứa nhiều bản ghi dữ liệu giao dịch. Mỗi bản ghi gồm các thông tin sau:

SERVICE_PK: Mã chỉ tiêu
 F_VALUE: giá trị
 F_VALUE_MONTH: giá trị lũy kế
 PRD_ID: ngày định dạng(yyyyMMdd)
 DEP_ID: mã thị trường
 UNIT_ID: đơn vị tính

3.4 Kết quả chương trình

3.4.1 Trung tâm GPCNTT Viettel

3.4.1.1 Biểu đồ lợi nhuận tháng theo doanh thu tài chính

3.4.1.1.1 Hướng dẫn xem biểu đồ



Hình 3.2 Hình vẽ vào chức năng xem biểu đồ

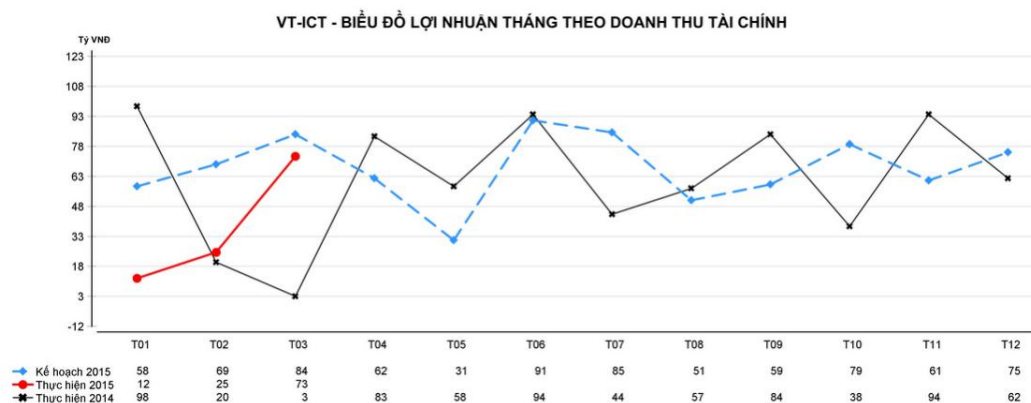


Hình 3.3 Hình vẽ màn hình tìm kiếm



Hình 3.4 Hình vẽ vào chức năng xem biểu đồ

Khi đó sẽ hiển thị Biểu đồ lợi nhuận tháng theo doanh thu tài chính năm 2015 theo số liệu đã nhập như sau:



Hình 3.5 Hình vẽ kết quả xem biểu đồ

Chú ý:

- + Dữ liệu năm hiện tại chỉ hiển thị tới tháng n-1. Trong đó: N là tháng hiện tại.
- + Người dùng có thể xem biểu đồ dưới dạng excel hoặc pdf bằng cách ấn nút



3.4.1.2 Biểu đồ lợi nhuận tháng theo tổng doanh thu

3.4.1.2.1 Hướng dẫn xem biểu đồ



Hình 3.6 Hình vẽ vào chức năng xem biểu đồ

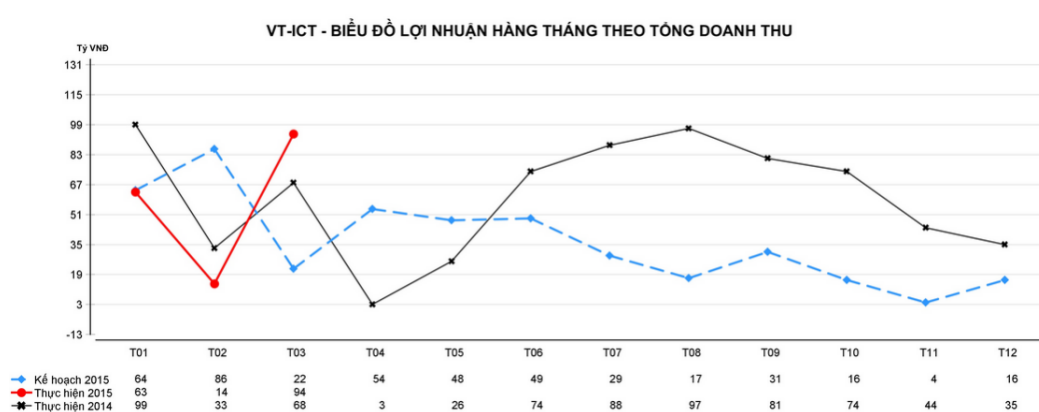


Hình 3.7 Hình vẽ vào chức năng xem biểu đồ



Hình 3.8 Hình vẽ vào chức năng xem biểu đồ

Khi đó sẽ hiển thị Biểu đồ lợi nhuận tháng theo tổng doanh thu năm 2015 theo số liệu đã nhập như sau:



Hình 3.9 Hình vẽ kết quả chức năng xem biểu đồ

Chú ý:

+ Dữ liệu năm hiện tại chỉ hiển thị tới tháng n-1. Trong đó: N là tháng hiện tại.

+ Người dùng có thể xem biểu đồ dưới dạng excel hoặc pdf bằng cách ấn nút



Excel

hoặc



PDF

3.4.1.3 Biểu đồ tiến độ launching sản phẩm đại trà/lãi

3.4.1.3.1 Hướng dẫn xem biểu đồ



Hình 3.10 Hình vẽ vào chức năng xem biểu đồ

Khi đó sẽ hiển thị màn hình biểu đồ như sau với năm là mặc định là năm hiện tại, Đơn vị mặc định tên group VTICT:

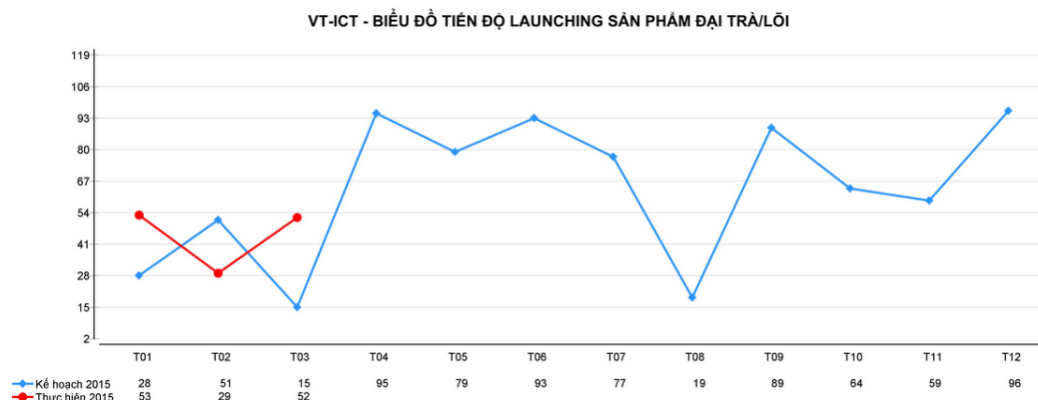


Hình 3.11 Hình vẽ vào chức năng xem biểu đồ



Hình 3.12 Hình vẽ vào chức năng xem biểu đồ

Khi đó sẽ hiển thị Biểu đồ tiến độ launching sản phẩm đại trà/lãi năm 2015 theo số liệu đã nhập như sau:



Hình 3.13 Hình vẽ kết quả chức năng xem biểu đồ

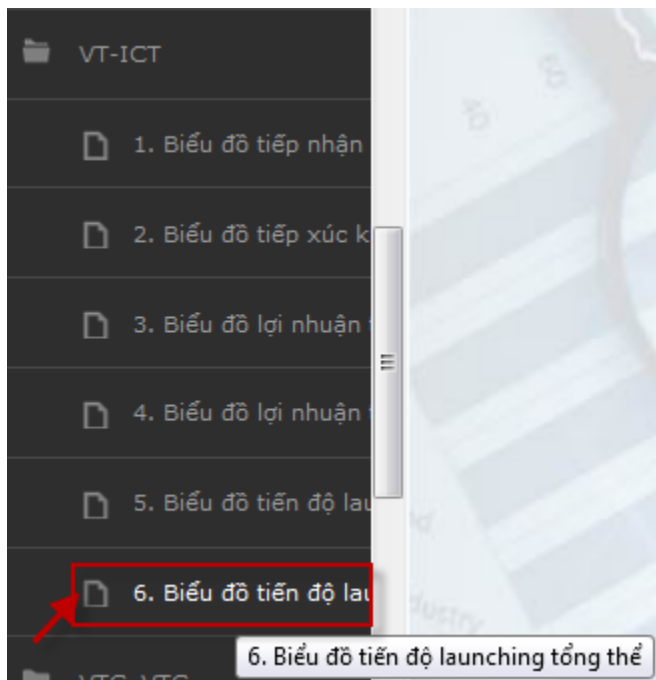
Chú ý:

- + Dữ liệu năm hiện tại chỉ hiển thị tới tháng n-1. Trong đó: N là tháng hiện tại.
- + Người dùng có thể xem biểu đồ dưới dạng excel hoặc pdf bằng cách ấn nút



3.4.1.4 Biểu đồ tiến độ launching tổng thể

3.4.1.4.1 Hướng dẫn xem biểu đồ



Hình 3.14 Hình vẽ vào chức năng xem biểu đồ

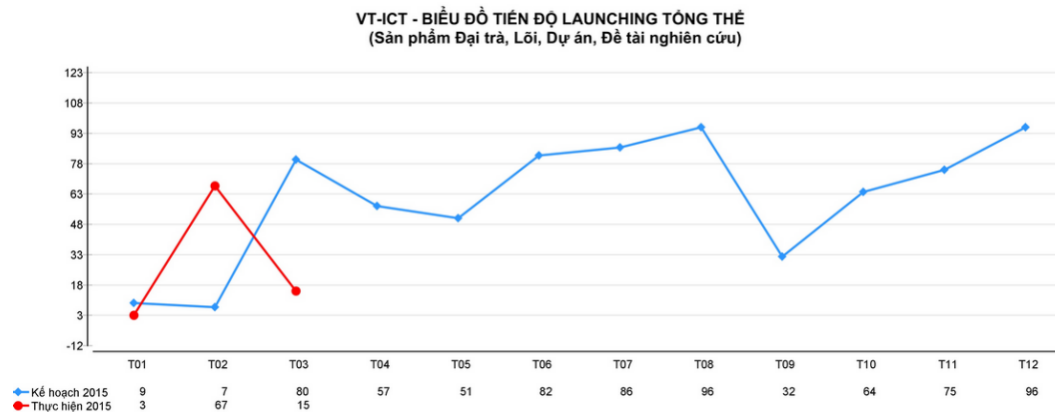


Hình 3.15 Hình vẽ vào chức năng xem biểu đồ



Hình 3.16 Hình vẽ vào chức năng xem biểu đồ

Khi đó sẽ hiển thị Biểu đồ tiến độ launching tổng thể (Sản phẩm Đại trà, Lỗi, Dự án, Đề tài nghiên cứu) năm 2015 theo số liệu đã nhập như sau:



Hình 3.17 Hình vẽ kết quả xem biểu đồ

Chú ý:

- + Dữ liệu năm hiện tại chỉ hiển thị tới tháng n-1. Trong đó: N là tháng hiện tại.
- + Người dùng có thể xem biểu đồ dưới dạng excel hoặc pdf bằng cách ấn nút



Excel

hoặc



PDF

3.5 Đánh giá chương trình

3.6 Kết luận chương

KẾT LUẬN

Trong thời đại hiện nay, cùng với sự bùng nổ về dữ liệu, khối lượng dữ liệu của các doanh nghiệp là vô cùng lớn. Việc phân tích, xử lý, khai thác dữ liệu Big data trong thời gian thực để đưa ra kết quả nhanh chóng, chính xác hỗ trợ các nhà quản lý đưa ra các quyết định kinh doanh kịp thời là hết sức quan trọng, tạo ra sức mạnh cạnh tranh rất lớn cho doanh nghiệp, đặc biệt là các doanh nghiệp viễn thông. Xuất phát từ bản thân tác giả là kỹ sư giải pháp phần mềm tại Tập đoàn Viettel, tác giả đã lựa chọn nghiên cứu các công nghệ Big data và ứng dụng để xây dựng hệ thống xử lý số liệu kinh doanh của tập đoàn Viettel để thực hiện luận văn của mình.

Với tác giả thì công nghệ Big data vẫn là một công nghệ mới, việc nghiên cứu trong một thời gian ngắn nên vẫn chưa khám phá lĩnh vực hết công nghệ này. Tuy nhiên qua quá trình nghiên cứu luận văn, tác giả đã thu được một số kết quả cũng như nhận thấy một số hạn chế như sau:

1. Kết quả đạt được

Về mặt lý thuyết, tác giả đã có những nghiên cứu về Big data, các công nghệ Big data trên thế giới, hiểu sâu về công nghệ Apache Hadoop.

Về mặt thực nghiệm, tác giả đã xây dựng được hệ thống xử lý số liệu kinh doanh của Tập đoàn Viettel dựa trên công nghệ Apache Hadoop. Tác giả cũng đã có những phân tích, đánh giá được kết quả thực nghiệm

2. Hạn chế

Kết quả mới được thực hiện trên bộ dữ liệu còn chưa đủ lớn (chỉ với dữ liệu thử nghiệm của 22 tháng từ năm 2014 đến năm 2016), mô hình phân tích còn đơn giản, mới chỉ tập trung vào dữ liệu có cấu trúc. Ngoài ra, do thời gian thực hiện luận văn có hạn nên tác giả chưa nghiên cứu để sử dụng các thuật toán học máy để đánh giá xu thế và dự báo kết quả kinh doanh trong tương lai.

3. Hướng phát triển

Trong thời gian tới, tác giả sẽ tiếp tục ứng dụng rộng rãi hệ thống cho các bộ dữ liệu thật, không có cấu trúc, tiếp tục nghiên cứu các thuật toán học máy để đánh giá xu thế cho việc phân tích dự báo kết quả trong tương lai.

TÀI LIỆU THAM KHẢO

Tài liệu Tiếng Việt

- [1] Tập san tin học quản lý Tập 03, số 1&2, 2014, 53-73 (2014). *Bigdata bức tranh toàn cảnh*, Khoa Hệ Thống Thông Tin Kinh Doanh – ĐH Kinh Tế HCM.

Tài liệu Tiếng Anh

- [2] O'Reilly Media Team (2012), *Big Data Now* (2012 Edition), O'Reilly Media.
- [3] Mike Barlow (2013), *Real-Time Big Data Analytics* (2013 Edition), O'Reilly Media.
- [4] Tom White (2012), *Hadoop: The Definitive Guide, 3rd Edition* (2012 Edition), O'Reilly Media.
- [5] Alex Holmes (2012), *Hadoop in Practice* (2012 Edition), Manning.