**CÔNG TY TNHH HỆ THỐNG THÔNG TIN FPT**

|  |  |
| --- | --- |
| **DỰ ÁN** **XÂY DỰNG CHÂN DUNG KHÁCH HÀNG**  **VÀ MÔ HÌNH AI PHÂN TÍCH DỮ LIỆU**  **CHO NGÂN HÀNG TMCP SÀI GÒN THƯƠNG TÍN**  **TÀI LIỆU THIẾT KẾ KIẾN TRÚC HẠ TẦNG** | |
| **Mã hiệu tài liệu** | CINS-TKKTHT |
| **Phiên bản** | 1.0 |
| **Ngày cập nhật** | 17/1/2022 |

Hà Nội, 11/2022

**Mục Lục**

[I. TỔNG QUAN 5](#_Toc120031480)

[I.1. Phạm vi tài liệu 5](#_Toc120031481)

[I.2. Thuật ngữ 5](#_Toc120031482)

[II. TIÊU CHÍ ĐÁNH GIÁ DỰ ÁN 6](#_Toc120031483)

[III. XÂY DỰNG KIẾN TRÚC HỆ THỐNG 7](#_Toc120031484)

[III.1. Kiến trúc giải pháp quản trị dữ liệu 7](#_Toc120031485)

[III.2. Kiến trúc triển khai và sizing chi tiết 8](#_Toc120031486)

[III.2.1. Môi trường DEV/TEST 8](#_Toc120031487)

[III.2.2. Môi trường UAT 11](#_Toc120031488)

[III.2.3. Môi trường PROD 13](#_Toc120031489)

[III.3. Mô hình kiến trúc 16](#_Toc120031490)

[III.3.1. Mô hình kiến trúc vật lý 16](#_Toc120031491)

[III.3.2. Mô hình kiến trúc luận lý 19](#_Toc120031492)

[III.4. Các thông tin kết nối 20](#_Toc120031493)

[IV. VẬN HÀNH HỆ THỐNG 22](#_Toc120031494)

[IV.1. Loging 22](#_Toc120031495)

[IV.2. Exception Handling 23](#_Toc120031496)

[IV.3. Security 25](#_Toc120031497)

[IV.4. Monitoring 26](#_Toc120031498)

[IV.5. Scalability 27](#_Toc120031499)

[IV.6. Availability 27](#_Toc120031500)

[IV.7. Backup and Restore 28](#_Toc120031501)

[IV.8. Archiving 28](#_Toc120031502)

[V. PHỤ LỤC 29](#_Toc120031503)

**Lịch sử thay đổi**

|  |  |
| --- | --- |
| **Phiên bản** | **Thay đổi** |
| 0.1 | Khởi tạo |
| 0.2 | Cập nhật mô hình vật lý, mô hình logic |
| 0.3 | Cập nhật mô hình luận lý |
| 0.4 | Cập nhật mô hình luận lý, mô tả các thành phần mới |
| 0.5 | Cập nhật thông số các máy chủ mới, cập nhật mô hình |
| 0.6 | Cập nhât tiêu chí đánh giá dự án, cập nhật thông tin về máy chủ kết nối GGAPI và Sharing làm giàu dữ liệu |
| 0.7 | Cập nhật mô hình triển khai tại môi trường Production |
| 1.0 | Release |

**TRANG KÝ**

**XÁC NHẬN CỦA FIS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Vai trò** | **Tên và chức vụ** | **Chữ ký** | **Ngày** |
| Người tạo | Lê Minh Nguyên | ............................... | ......................... |
| Người xem xét | Nguyễn Đức Trung Dũng | ............................... | ......................... |
|  |  |  |  |

**XÁC NHẬN CỦA SACOMBANK**

|  |  |  |  |
| --- | --- | --- | --- |
| **Vai trò** | **Tên và chức vụ** | **Chữ ký** | **Ngày** |
| Người xem xét | ……………………. | ............................... | ......................... |
| Người xét duyệt | ............................... | ............................... | ......................... |

1. TỔNG QUAN
   1. Phạm vi tài liệu

Tài liệu này mô tả về kiến trúc hệ thống được xây dựng cho Sacombank để phục vụ dự án xây dựng chân dung khách hàng và mô hình AI phân tích dữ liệu.

* 1. Thuật ngữ

|  |  |
| --- | --- |
| **THUẬT NGỮ** | **MÔ TẢ** |
| FPT-IS | FPT Information System |
| BDA | Big Data and Analytics |
| DEV | Developer |
| DRC | Disaster Recovery Center |
| DWH | Data Ware House |
| ERP | Enterprise Resource Planning |
| ETL | Extraction, Transformation and Loading |
| IT | Information Technology |
| ITD | IT department |
| LDAP | Lightweight Directory Access Protocol |
| NFS | Network File System |
| ODI | Oracle Data Integrator |
| OAS | Oracle Analytics Server |
| PROD | Production |
| SAN | Storage Area Network |
| SSL | Secure Socket Layer |
| UAT | User Acceptance Testing |
| VPN | Virtual Private Network |

1. TIÊU CHÍ ĐÁNH GIÁ DỰ ÁN
2. Các thành phần ứng dụng phân tích sẽ được thiết kế động bộ chung thiết bị:

Ứng dụng JupyterHub được triển khai trên hệ thống Openshift của Sacombank, các máy chủ ODI và OAS sử dụng tài nguyên đã có sẵn, máy chủ File Server và GoogleGW là các máy ảo dược Sacombank cung cấp.

1. Hệ thống được thiết kế tuân thủ theo mô hình dữ liệu luận lý của Sacombank.
2. Dữ liệu được lấy từ kho dữ liệu thông qua lớp truy cập và phân phối dữ liệu, đảm bảo mọi nguồn đều được đăng ký và xác nhận

* Các tài khoản truy vẫn dữ liệu đều được cung cấp thông qua PYC của Sacombank
* Các Schema được xin cấp tạo mới thông qua phòng quản trị dữ liệu
* Các bảng cần trích xuất hoặc mapping đều được đăng kí và thông qua và cấp quyền bởi Sacombank

1. Các ứng dụng của hệ thống được đặt gần dữ liệu để giảm thiểu sự di chuyển dữ liệu và trùng lặp dữ liệu:

* Các ứng dụng tạo job ETL và Dashboard, báo cáo (ODI, OAS) phụ thuộc theo thiết kế cũ của Sacombank
* Ứng dụng JupyterHub trên Openshift nằm trên lớp ứng dụng và được mở cổng truy vấn dữ liệu trên DWH

1. Các mức dịch vụ được thiết lập cho thành phần thăm dò và khám phá đều được thông qua từ các bên nghiệp vụ, được đồng ý bởi Sacombank
2. Tất cả các ứng dụng phân tích sẽ có quyền truy cập Read-only vào dữ liệu trong hệ sinh thái phân tích:

* Các tài khoản được cấp sẽ chỉ có quyền Read dữ liệu nguồn
* Các schema xin cấp mới sẽ có quyền đọc ghi dữ liệu, phục vụ cho việc chạy job ETL

1. Dựa trên yêu cầu của người dùng để lựa chọn các công cụ phân tích thích hợp. JupyterHub đáp ứng nhu cầu về môi trường lập trình Python cho Developer, ODI cho nhu cầu ETL, OAS xây dựng báo cáo, dashboard.
2. Tất cả quyền truy cập của người dùng sẽ được quản lý thông qua lớp truy cập dữ liệu, JupyterHub có khả năng tích hợp với các hệ thống xác thực bên ngoài (LDAP, OpenID…)
3. Quản lý hiệu suất sẽ được tách biệt để có hiệu suất tốt nhất, JupyterHub được quản lý hiệu suất bởi Openshift
4. Các trang dashboard tổng quan và báo cáo được làm theo các tiêu chí mà các phòng bạn nghiệp vụ đề xuất, cũng như theo yêu cầu của Sacombank
5. Hệ thống cho phép đồng bộ log để monitoring hiệu suất.
6. Dữ liệu dành cho xây dựng mô hình hay xây dựng dashboard hoàn toàn không hiển thị hoặc sử dụng thông tin nhận dạng cá nhân
7. Báo cáo đặc biệt và yêu cầu dữ liệu chạy thường xuyên phải được chuyển đổi thành báo cáo chuẩn. Trong phạm vi 12 tuần chưa có job chạy realtime, chỉ có job chạy daily. Các báo cáo sẽ thiết kế theo yêu cầu của nghiệp vụ.
8. Thành phần báo cáo được thiết kế để hỗ trợ các báo cáo tự động và phân tích nếu xảy ra: Chưa có yêu cầu tự động phân tích từ nghiệp vụ.
9. Các mô hình dự báo và phân tích được lập trình tối ưu hoá, đảm bảo hiệu suất cao.
10. Các job ETL/ELT sẽ được thực hiện trên môi trường ODI của Sacombank, được bàn giao lại sau khi golive hệ thống.
11. XÂY DỰNG KIẾN TRÚC HỆ THỐNG
    1. Kiến trúc giải pháp quản trị dữ liệu

Diagram

Description automatically generated

Hình 1. Kiến trúc giải pháp

Kiến trúc nền tảng khoa học dữ liệu bao gồm 4 thành phần chính, xuất phát từ data sources:

1. **Data sources:** bao gồm các nguồn dữ liệu chính:

* EDW: Kho dữ liệu
* FPT CDP: Nguồn dữ liệu làm giàu bổ sung thông tin tài chính giúp nâng cao độ chính xác mô hình

1. **Data Ingestion:**

Gồm công cụ thu nhận và tích hợp dữ liệu, Kết nối với các Data Sources, sử dụng công cụ Oracle Data Integrator (ODI) thu thập, định tính, xử lý và di chuyển dữ liệu theo lô (batch) để chuẩn bị lưu trữ trong Kho lưu trữ dữ liệu, sau đó được chia sẻ với các ứng dụng và hệ thống Phân tích / Truy cập sau này. Chia làm 3 chức năng nhỏ hơn là:

* Chuyển đổi dữ liệu: Áp dụng các luật nghiệp vụ (Business rules) để chuẩn hoá dữ liệu phù hợp với mô hình dữ liệu đích.
* Chất lượng dữ liệu: Bao gồm các quy tắc chuẩn hóa và chất lượng dữ liệu trên dữ liệu nguồn để đảm bảo rằng nó tuân thủ các tiêu chuẩn và được mọi người cần sử dụng hiểu rõ.
* Tải dữ liệu: Quá trình tải hoặc chèn dữ liệu vào kho lưu trữ đích (nguồn phân tích) và làm cho dữ liệu luôn sẵn sàng để sử dụng. Việc này có thể được lập kế hoạch để xử lý theo lịch định kỳ.

1. **Data Management & Storage**: Khối chức năng định ra các công cụ quản lý và lưu trữ dữ liệu, đây là phần cơ bản nhất của các kho dữ liệu hay hồ dữ liệu. Sử dụng storage có sẵn là Oracle Exadata. Được chia thành các vùng dữ liệu:

* Landing Zone: Nơi tập trung lưu trữ dữ liệu thô, được đồng bộ từ các hệ thống tác nghiệp, chủ yếu phục vụ yêu cầu báo cáo vận hành, off-load các tiến trình xử lý dữ liệu.
* Feature Store: Kho đặc trưng, được tổng hợp, xử lý sẵn sàng để sử dụng cho các bài toán mô hình.
* Datamart: Lưu dữ liệu tổng hợp định hướng khai thác, phục vụ cho việc khai thác dữ liệu dưới dạng các báo cáo biểu đồ.
* SMY - Data Summary: Dữ liệu quy hoạch theo các chủ đề Customer Insight, được tổng hợp.

1. **Data Driven Decision Making** bao gồm các công cụ khai thác dữ liệu:

* Công cụ tìm kiếm, trực quan hoá dữ liệu dưới dạng các Báo cáo, Dashboard (Sử dụng Oracle Analytic Server) sẽ giúp cho người dùng nghiệp vụ (business user) dễ dàng tiếp cận với dữ liệu và thông tin khai thác được.
* Công cụ khai thác, phân tích dữ liệu chuyên sâu (Sử dụng Oracle Data Science kết hợp với Oracle Analytic Server): đây là phần quan trọng khi đem đến các năng lực và ứng dụng AI vào phân tích dữ liệu, các kỹ thuật của khoa học dữ liệu để hiểu dữ liệu một cách sâu sắc, đa chiều. Khối chức năng này cung cấp các công cụ, các mô hình, thuật toán AI/ML để các nhà khoa học dữ liệu, kỹ sư dữ liệu tạo ra các mô hình khai thác dữ liệu một cách hiệu quả.
  1. Kiến trúc triển khai và sizing chi tiết
     1. Môi trường DEV/TEST
        1. Kiến trúc

Graphical user interface

Description automatically generated with low confidence

Hình 2. Kiến trúc triển khai Dev/Test

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Interface name** | **Source** | **Target** | **Integration Method** | **Description** |
| I-01 | Server Access Internet | File Server | NFS  TCP/UDP 111 (2 chiều) | Push dữ liệu FPT.CDP Data về File Server |
| I-02 | Server Access Internet | File Server | NFS  TCP/UDP 111 (2 chiều) | Transfer dữ liệu Google API đã xử lý về File Server |
| I-03 | ODI | Oracle Exadata | ODI Studio  TCP 1521/1594 | ETL dữ liệu nguồn, đẩy vào vùng thô |
| I-04 | ODI | Oracle Exadata | ODI Studio  TCP 1521/1594 | ETL dữ liệu thô sang kho ADS |
| I-05 | ODI | Oracle Exadata | ODI Studio  TCP 1521/1594 | ETL kho ADS tạo các bảng SMY phục vụ bài toán |
| I-06 | OAS | Oracle Exadata | Web-based  TCP 1521/1594 | OAS lấy dữ liệu từ các bảng SMY để tạo báo cáo, dashboard |
| I-07 | JupyterHub | Oracle Exadata | Web-based  TCP 1521/1594 | Jupyter sử dụng dữ liệu SMY training model |
| I-08 | JupyterHub | File Server | NFS  TCP/UDP 111 (2 chiều) | Trích xuất dữ liệu phi cấu trúc (Logs, csv, pdf..) |

Tầng Database Services:

* Oracle Exadata bao gồm các vùng dữ liệu sau:
* Vùng RAW (ADS) chứa dữ liệu thô ETL từ Data Source trên hệ thống Exadata
* Vùng Summary (SMY) dữ liệu theo các chủ đề, được ETL tổng hợp lại từ dữ liệu thô
* Vùng FeatureStore (FS) kho đặc trưng, được tổng hợp, xử lý sẵn sàng để sử dụng cho các bài toán mô hình.
* File Server: Phân vùng NFS, lưu trữ dữ liệu phi cấu trúc (logs file, csv, txt…) phục vụ cho việc trích xuất của JupyterHub

Tầng Analytic Services:

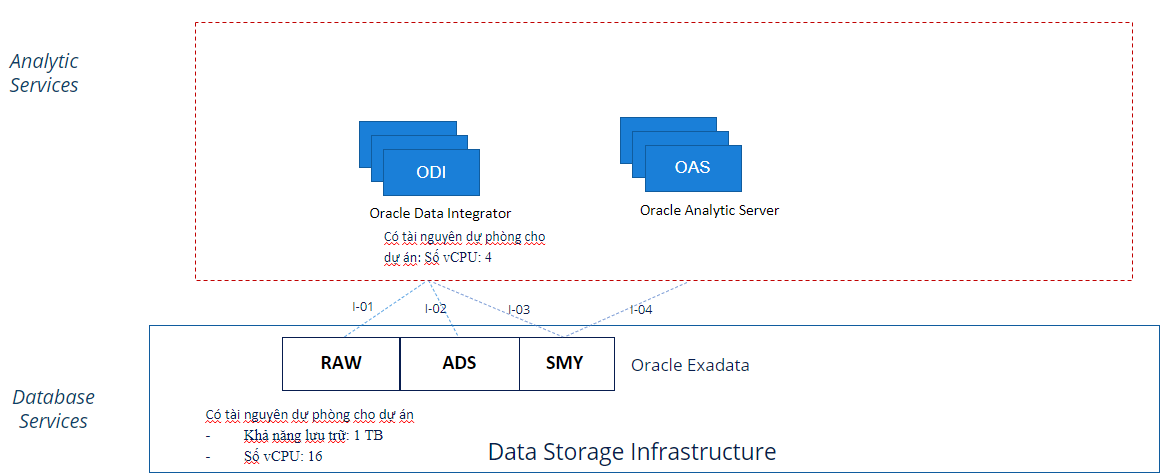
* Oracle Data Integrator (ODI): Chạy các Job ETL dữ liệu từ Datasource tới kho lưu trữ dữ liệu thô, từ đó tiếp tục transform dữ liệu phục vụ các bài toán phân tích.
* Oracle Analytic Server (OAS): Lấy dữ liệu từ kho dữ liệu phân tích phục vụ cho việc tạo các dashboard, báo cáo về Customer360.
* JupyterHub: Python Dev Platform, lấy dữ liệu từ kho ADS cho việc training model, các bài toán AL/ML. Ngoài ra sử dụng File Server để trích xuất dữ liệu phi cấu trúc.
* Internet Access Server: Kết nối tới máy chủ lưu dữ liệu FPT.CDP từ FIS, lấy dữ liệu về Sacombank, sử dụng để làm giàu dữ liệu, phục vụ các bài toán training model. Ngoài ra truy cập ra internet tới dịch vụ Google API để xử lý location chính xác hơn, từ đó có dữ liệu để xây dựng các dashboard theo vị trí địa lý

Tầng Data Science (Cloud): Jupyter Notebook được cung cấp trên môi trường OCI, phục vụ vụ các bài toán lớn, cần sử dụng nhiều tài nguyên

* + - 1. Sizing thiết kế hệ thống

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Tên máy chủ** | **Hệ điều hành** | **Địa Chỉ IP** | **Cấu hình** | **Ghi chú** |
| 1 | Oracle Data Integrator | Linux | 192.168.124.108 | 8 vCPU | On-Premise (Reserve tài nguyên sẵn có) |
| 2 | Oracle Analytic Server | Linux | 192.168.105.135 | 4 vCPU  32GB RAM | On-Premise  (Reserve tài nguyên sẵn có) |
| 3 | Jupyter Hub | Cài đặt trên Openshift | https://jupyter.apps.ocptest.sacombank.local | 33 vCPU  86GB RAM  420GB PVC | On-Premise  Cài mới  Internet Access qua Proxy |
| 4 | Oracle Exadata |  | 192.168.124.100 --> 110 (11 Ips) | "Khả năng lưu trữ: 1TB  - Kho ADS ( Schema: ADS)  - Kho SMY (Schema: SMY)  - Kho FeatStore (Schema: FeatStore)  - Customer Segmentation Sandbox (Schema: sb\_segment\_cn)  - Customer Reactivation Sandbox (Schema: sb\_reactivate\_cn)  - Cross and Up Sell Sandbox (Schema: sb\_crossup\_cn)"  Số vCPU: 16 | On-Premise |
| 5 | File storage | NFS | 192.168.117.209  /data-fis/data-log/ | Khả năng lưu trữ: 2TB | On-Premise |
| 6 | Server Internet Access | Windows 10 | Sacombank cung cấp môi trường | Reserve 4CPU, 8GB RAM  Khả năng lưu trữ 256GB  Access URL:  <https://docs.anaconda.com/anaconda/install/windows/>  <https://pypi.org>  [https://maps.googleapis.com/maps/api/geocode/](https://maps.googleapis.com/maps/api/geocode/%20%20)  <https://maps.googleapis.com/maps/api/>  https://lakedpa.apps.xplat.fis.com.vn | On-prem  Internet Access  Anaconda for Windows |

* + 1. Môi trường UAT
       1. Kiến trúc



Hình 3. Kiến trúc triển khai UAT

*Cùng hệ thống Oracle Exadata với môi trường DEV/TEST*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Interface name** | **Source** | **Target** | **Integration Method** | **Description** |
| I-01 | ODI | Oracle Exadata | ODI Studio  TCP 1521/1594 | Chạy job ETL migrate từ dev/test dữ liệu nguồn, đẩy vào vùng thô |
| I-02 | ODI | Oracle Exadata | ODI Studio  TCP 1521/1594 | Chạy job ETL migrate từ dev/test dữ liệu thô sang kho ADS |
| I-03 | ODI | Oracle Exadata | ODI Studio  TCP 1521/1594 | Chayj job ETL migrate từ dev/test kho ADS tạo các bảng SMY phục vụ bài toán |
| I-04 | OAS | Oracle Exadata | Web-based  TCP 1521/1594 | OAS lấy dữ liệu từ các bảng SMY để tạo báo cáo, dashboard |

Tầng Database Services:

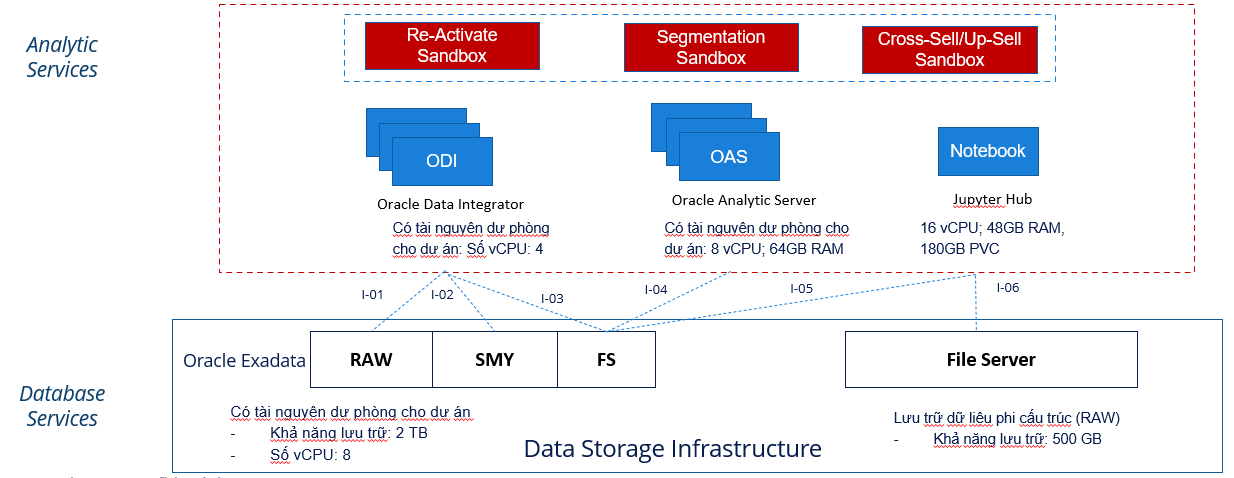
* Oracle Exadata:Cùng hệ thống Oracle Exadata với môi trường DEV/TEST

Tầng Analytic Services:

* Oracle Data Integrator (ODI): Chạy các Job ETL dữ liệu (được migrate từ môi trường DEV/TEST).
* Oracle Analytic Server (OAS): Migrate các báo cáo, dashboard từ môi trường DEV/TEST
  + - 1. Sizing thiết kế hệ thống

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Tên máy chủ** | **Hệ điều hành** | **Địa chỉ IP** | **Cấu hình** | **Ghi chú** |
| 1 | Oracle Data Integrator | Linux |  | 4 vCPU | On-Premise |
| 2 | Oracle Analytic Server | Linux |  | 8 vCPU  64GB RAM | On-Premise |
| 3 | Oracle Exadata |  | 192.168.124.100 --> 110 (11 Ips) | "Khả năng lưu trữ: 1TB  - Kho ADS ( Schema: ADS)  - Kho SMY (Schema: SMY)  - Kho FeatStore (Schema: FeatStore)"  Số vCPU: 16 | On-Premise |

* + 1. Môi trường PROD
       1. Kiến trúc



Hình 4. Kiến trúc triển khai Production

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Interface name** | **Source** | **Target** | **Integration Method** | **Description** |
| I-01 | ODI | Oracle Exadata | ODI Studio  TCP 1521/1594 | Chạy job ETL (migrate từ UAT) dữ liệu nguồn, đẩy vào vùng thô |
| I-02 | ODI | Oracle Exadata | ODI Studio  TCP 1521/1594 | Chạy job ETL (migrate từ UAT) dữ liệu thô sang kho ADS |
| I-03 | ODI | Oracle Exadata | ODI Studio  TCP 1521/1594 | Chạy job ETL (migrate từ UAT) kho ADS tạo các bảng SMY phục vụ bài toán |
| I-04 | OAS | Oracle Exadata | Web-based  TCP 1521/1594 | OAS lấy dữ liệu từ các bảng SMY để tạo báo cáo, dashboard |
| I-05 | JupyterHub | Oracle Exadata | Web-based  TCP 1521/1594 | Jupyter sử dụng dữ liệu SMY training model |
| I-06 | JupyterHub | File Server | NFS  TCP/UDP 111 (2 chiều) | Trích xuất dữ liệu phi cấu trúc (Logs, csv, pdf..) |

Tầng Database Services:

* Oracle Exadata bao gồm các vùng dữ liệu sau:
  + Vùng RAW (ADS) chứa dữ liệu thô ETL từ Data Source trên hệ thống Exadata
  + Vùng Summary (SMY) dữ liệu theo các chủ đề, được ETL tổng hợp lại từ dữ liệu thô, các job ETL được migrate từ môi trường DEV/TEST
  + Vùng FeatureStore (FS) kho đặc trưng, được tổng hợp, xử lý sẵn sàng để sử dụng cho các bài toán mô hình, các job ETL được migrate từ môi trường DEV/TEST
* File Server: Phân vùng NFS, lưu trữ dữ liệu phi cấu trúc (logs file, csv, txt…) phục vụ cho việc trích xuất của JupyterHub

Tầng Analytic Services:

* Oracle Data Integrator (ODI): Migrate các job từ môi trường UAT, xử lý dữ liệu cho môi trường PROD
* Oracle Analytic Server (OAS): Lấy dữ liệu từ kho dữ liệu phân tích phục vụ cho việc tạo các dashboard, báo cáo về Customer360. Migrate các dashboard, báo cáo từ UAT
* JupyterHub: Python Dev Platform, import các notebook file (.ipynb) từ Jupyter môi trường Dev/Test, chạy các bài toán trên môi trường PROD
  + - 1. Sizing thiết kế hệ thống

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Tên máy chủ** | **Hệ điều hành** | **Địa chỉ IP** | **Cấu hình** | **Ghi chú** |
| 1 | Oracle Data Integrator | Linux |  | 4 vCPU | On-Premise |
| 2 | Oracle Analytic Server | Linux |  | 8 vCPU  64GB RAM | On-Premise |
| 3 | Jupyter Lab | Cài đặt trên Openshift |  | 16 vCPU  48GB RAM  180GB PVC | On-Premise  Cài mới  Internet Access |
| 4 | Oracle Exadata | IP: 192.168.184.1 --> 11 (11 IPs) |  | "Khả năng lưu trữ: 2TB  - Kho ADS ( Schema: ADS)  - Kho SMY (Schema: SMY)  - Kho FeatStore (Schema: FeatStore)  - Customer Segmentation Sandbox (Schema: sb\_segment)  - Customer Reactivation Sandbox (Schema: sb\_reactivate)  - Cross and Up Sell Sandbox (Schema: sb\_crossup)"  Số vCPU: 8 | On-Premise |
| 5 | File storage | NFS (SAN) |  | Khả năng lưu trữ: 500GB | On-Premise |

* 1. Mô hình kiến trúc
     1. Mô hình kiến trúc vật lý



Hình 5. Mô hình kiến trúc vật lý

Hệ thống bao gồm ứng dụng JupyterHub được cài mới trên nền tảng Openshift tại môi trường DEV/UAT và môi trường PROD.

Máy chủ Internet Access trên môi trường DEV được Sacombank cung cấp sử dụng vào việc trích xuất dữ liệu location của GoogleAPI, đồng thời lấy dữ liệu CDP làm giàu từ môi trường của FPT, lưu về môi trường DEV của Sacombank

Các máy chủ còn lại đã có sẵn, các cấu hình và kết nối cũ không cần chỉnh sửa thêm

Trong đó:

Môi trường DEV/TEST:

* 01 JupyterHub Server (Cài mới) trên môi trường Openshift

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Role** | **Pods** | **CPU/pod** | **RAM/pod (GB)** | **PVC/pod (GB)** | **total CPU** | **total RAM (GB)** | **total PVC (GB)** |
| Service Pod | 7 | 1 | 2 | 60 (Chỉ cần cho Postgres lưu cấu hình) | 7 | 14 | 60 |
| Normal User (Dự kiến 6 pod) | 6 | 4 | 12 | 60 | 24 | 72 | 360 |
| Total |  |  |  |  | 33 | 86 | 420 |

* Cụm máy chủ ODI Server (Có sẵn)

Cấu hình: 8 vCPU

* Cụm máy chủ OAS Server (Có sẵn)

Cấu hình: 4 vCPU, 32GB RAM

* Máy chủ Internet Access: (Xin cấp mới)

Cấu hình: 4 vCPU, 8GB RAM

* Máy chủ ExaData (Có sẵn)

Cấu hình: Lưu trữ 2TB, 16 vCPU

Môi trường UAT:

* Cụm máy chủ ODI Server (Có sẵn)

Cấu hình: 4 vCPU

* Cụm máy chủ OAS Server (Có sẵn)

Cấu hình: 8 vCPU, 64GB RAM

Môi trường PROD:

* 01 JupyterHub Server (Cài mới) trên môi trường openshift

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Role** | **Pods** | **CPU/pod** | **RAM/pod (GB)** | **PVC/pod (GB)** | **total CPU** | **total RAM (GB)** | **total PVC (GB)** |
| Service Pod | 7 | 1 | 2 | 60 (Chỉ cần cho Postgres lưu cấu hình) | 7 | 14 | 60 |
| Normal User (Dự kiến 6 pod) | 2 | 4 | 16 | 60 | 8 | 32 | 120 |
| Total |  |  |  |  | 15 | 48 | 180 |

* Cụm máy chủ ODI Server (Có sẵn)

Cấu hình: 4 vCPU

* Cụm máy chủ OAS Server (Có sẵn)

Cấu hình: 8 vCPU, 64GB RAM

* Máy chủ ExaData (Có sẵn)

Cấu hình: Lưu trữ 2TB, 8 vCPU

File Server (Storage):

* Khả năng lưu trữ DEV/UAT: 1TB
* Khả năng lưu trữ PROD: 500GB
  + 1. Mô hình kiến trúc luận lý



Hình 6. Mô hình kiến trúc luận lý

* Internet Area: Gồm hệ thống Oracle Cloud Infrastructure, sử dụng Oracle Data Science để xử lý các bài toán training model khi tài nguyên của JupyterHub onpremise không đủ. Sử dụng dữ liệu từ các Feature Store để training. (Chỉ deploy tại môi trường DEV)
* Sacombank Internal:
  + Application Zone: gồm các ứng dụng ODI, OAS, JupyterHub
  + Database Zone: gồm các kho dữ liệu phân vùng nằm trong Exadata và File Server
  + DMZ Zone: máy chủ Internet Access (Chỉ deploy tại môi trường DEV)

**Luồng xử lý:**

* Data Engineer truy cập vào Access Gateway để sử dụng ODI Server, tạo các job ETL, xử lý dữ liệu trong Exadata.
* Data Engineer truy cập vào Access Gateway remote tới máy chủ Internet Access lấy dữ liệu CDP từ hệ thống lưu trũ của FPT cung cấp, dữ liệu sau đó được đẩy tới File Server (NFS), phục vụ làm giàu dữ liệu.

Từ máy chủ này cũng kết nối sử dụng Cloud API được cung cấp bởi Google để làm chuẩn dữ liệu Location, sau đó nội bộ Sacombank sẽ lấy dữ liệu đã được xử lý đưa về File Server (NFS), phục vụ việc phân tích, tạo các dashboard, báo cáo theo vị trí địa lý.

* Data Science truy cập vào Gateway, sử dụng OAS Server tích hợp với dữ liệu tại Exadata, tạo các Dashboard và báo cáo về Customer 360
* Data Science truy cập vào Gateway, sử dụng Jupyter làm môi trường lập trình Python, trích xuất dữ liệu từ các kho tại Exadata để training model, lưu các model trên Exadata. Các dữ liệu phi cấu trúc sẽ được trích xuất tại phân vùng NFS (File Server)
* Các job training lớn sẽ được chạy trên Oracle Data Science môi trường cloud, các dữ liệu được lấy từ Feature Store trong kho DWH Exadata
  1. Các thông tin kết nối

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Nguồn** | **Đích** | **Port** | **Ghi chú** |
| 1 | JupyterHub server | Oracle Exadata (DEV/UAT)  IP: 192.168.124.100 --> 110 (11 Ips) | TCP 1521 | Kết nối từ Jupyter tới database trên exadata môi trường DEV/UAT |
| 2 | JupyterHub server | Oracle Exadata (PROD)  IP: 192.168.184.1 --> 11 (11 IPs) | TCP 1521 | Kết nối từ Jupyter tới database trên exadata môi trường PROD |
| 3 | JupyterHub server | File Storage (DEV/UAT) IP: 192.168.117.209 | NFS | Kết nối từ Jupyter file storage để đọc ghi dữ liệu phi cấu trúc môi trường DEV/UAT |
| 4 | JupyterHub server | File Storage (PROD) IP: | NFS | Kết nối từ Jupyter file storage để đọc ghi dữ liệu phi cấu trúc môi trường PROD |
| 5 | Client | JupyterHub Server  (DEV/UAT) | HTTPS 443 | Kết nối từ người dùng tới Jupyter Hub môi trường DEV/UAT |
| 6 | Client | JupyterHub Server  (PROD) | HTTPS 443 | Kết nối từ người dùng tới Jupyter Hub môi trường PROD |
| 8 | Client | Internet Access Server  (DEV/UAT) | 3389 RDP | Kết nối từ người dùng tới Google API để xử lý dữ liệu bản đồ môi trường DEV/UAT  Push/Get dữ liệu FPT.CDP |
| 9 | Internet Access Server | Internet  (DEV/UAT) | HTTPS 443 | Truy cập vào các địa chỉ  https://lakedpa.apps.xplat.fis.com.vn  https://maps.googleapis.com/maps/api  https://maps.googleapis.com/maps/api/geocode/ |

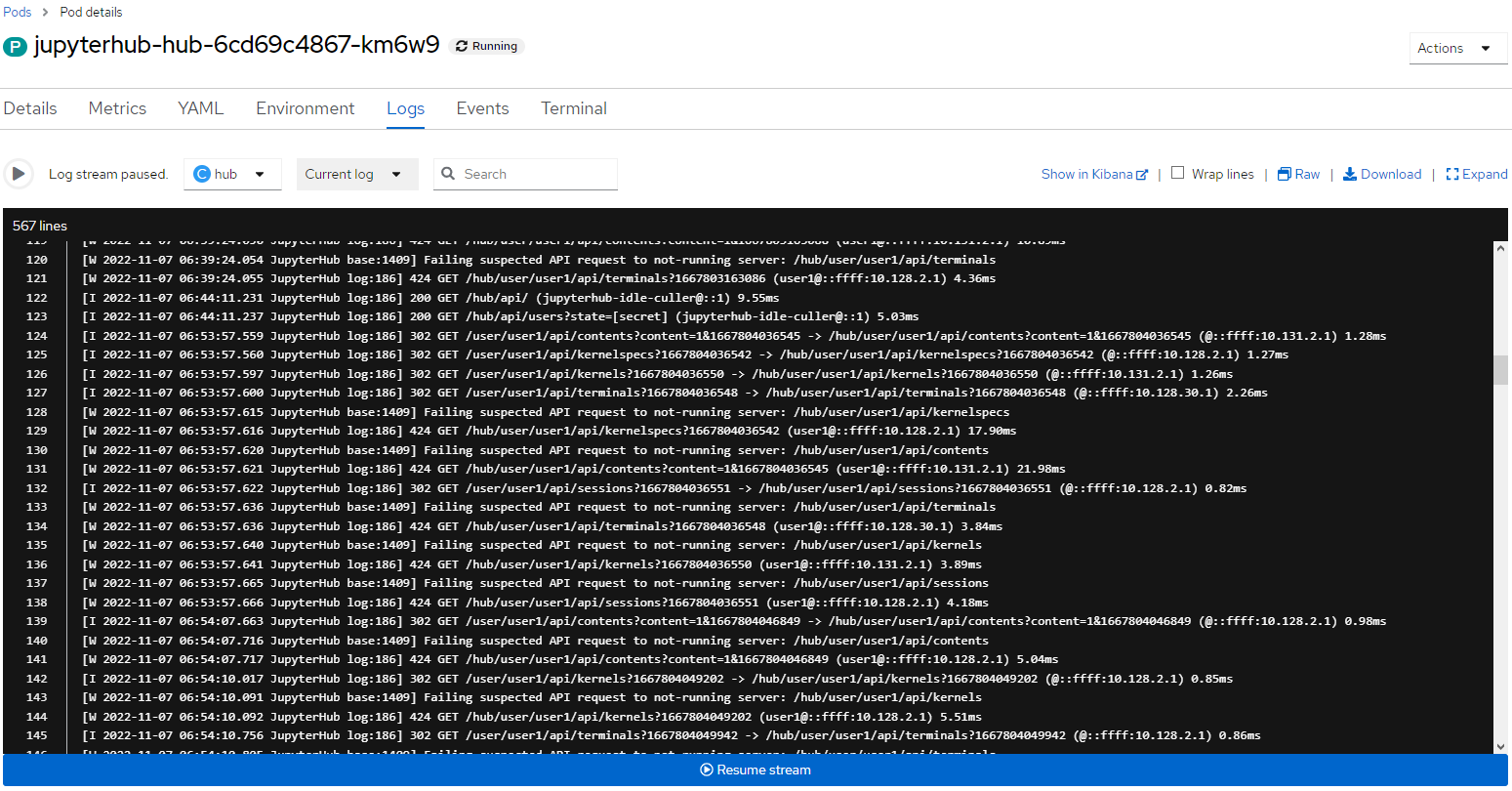
1. VẬN HÀNH HỆ THỐNG
   1. Loging

Quản lí logging cho hệ thống JupyterHub:

Log của hệ thống bao gồm log của Hub, User, Proxy. Được quản lí trên giao diện của Openshift

* + Hub: Service chạy chính của Jupyter, quản trị các cấu hình và spawn ra các user
  + Proxy: Service chạy dịch vụ HTTP/HTTPS cho JupyterHub
  + User: Các User pod do Hub spawn ra khi user đăng nhập vào hệ thống

Chọn các pod muốn theo dõi log -> Viewlogs



Hình 7. Log service JupyterHub

Text

Description automatically generated

Hình 8. Log user JupyterHub

A screenshot of a computer

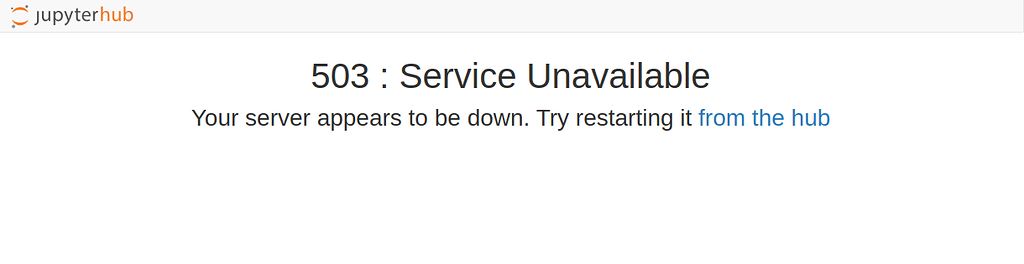
Description automatically generated

Hình 9. Log service HubProxy

* 1. Exception Handling

Các lỗi thường gặp trong quá trình vận hành hệ thống JupyterHub:

* + Lỗi 503 Service Unavailable: Lỗi thường gặp khi service Hub có lỗi xảy ra, hoặc cấu hình sai trong file cấu hình
    - Xử lý: Kiểm tra log của pod Hub service, kiểm tra lại các cấu hình trong file values.yalm sau đó delete pod cũ và tạo pod mới cho Hub service



Hình 10. Lỗi 503 Service Unavailable

* + Lỗi User pod không thể tạo do thiếu tài nguyên (Insufficent memory)
    - Xử lý: Tăng tài nguyên cho hệ thống Openshift bằng việc thêm các Worker Node vào trong Opensfhit cluster. Hoặc cấu hình giảm tài nguyên sử dụng của các User pod trong file cấu hình values.yalm:

*{singleuser:*

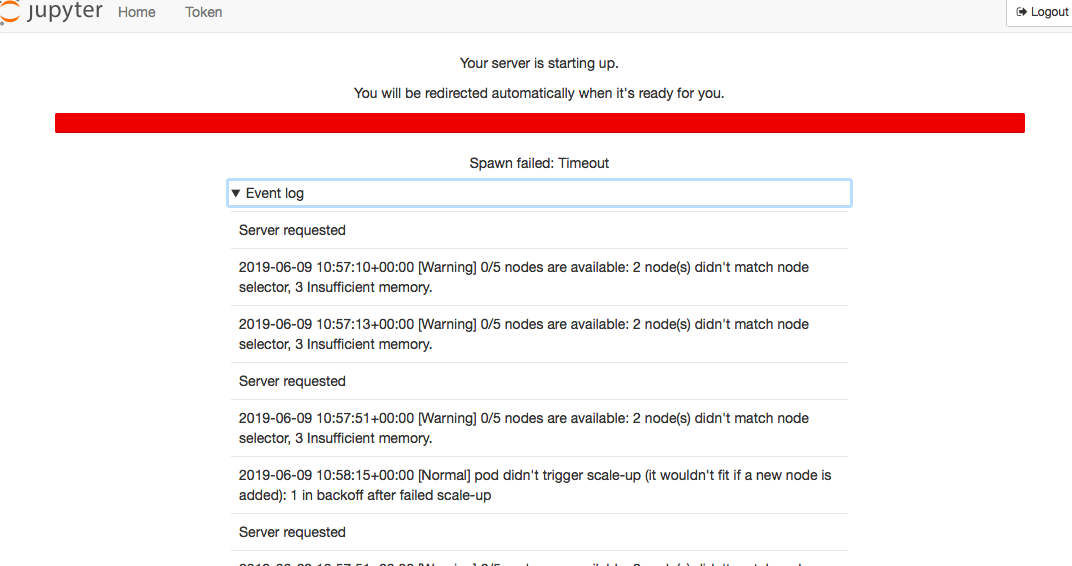
*memory:*

*limit:*

*guarantee: }*

Trong đó:

* + guarantee là số lượng tài nguyên tối thiểu User luôn có thể sử dụng
  + limit là số lượng tài nguyên tối đa User có thể sử dụng



Hình 11. Lỗi thiếu tài nguyên

* 1. Security

Hệ thống JupyterHub triển khai trên Openshift cluster được cấu hình HTTPS, bảo đảm an toàn bảo mật khi truy cập từ bên ngoài thông qua giao diện người dùng Web-based

Graphical user interface, text, application, email

Description automatically generated

Hình 12. Cấu hình HTTPS 443

Authentication:

Hệ thống Jupyter cho phép cấu hình xác thực, tích hợp với nhiều hệ thống quản trị người dùng từ bên ngoài bao gồm:

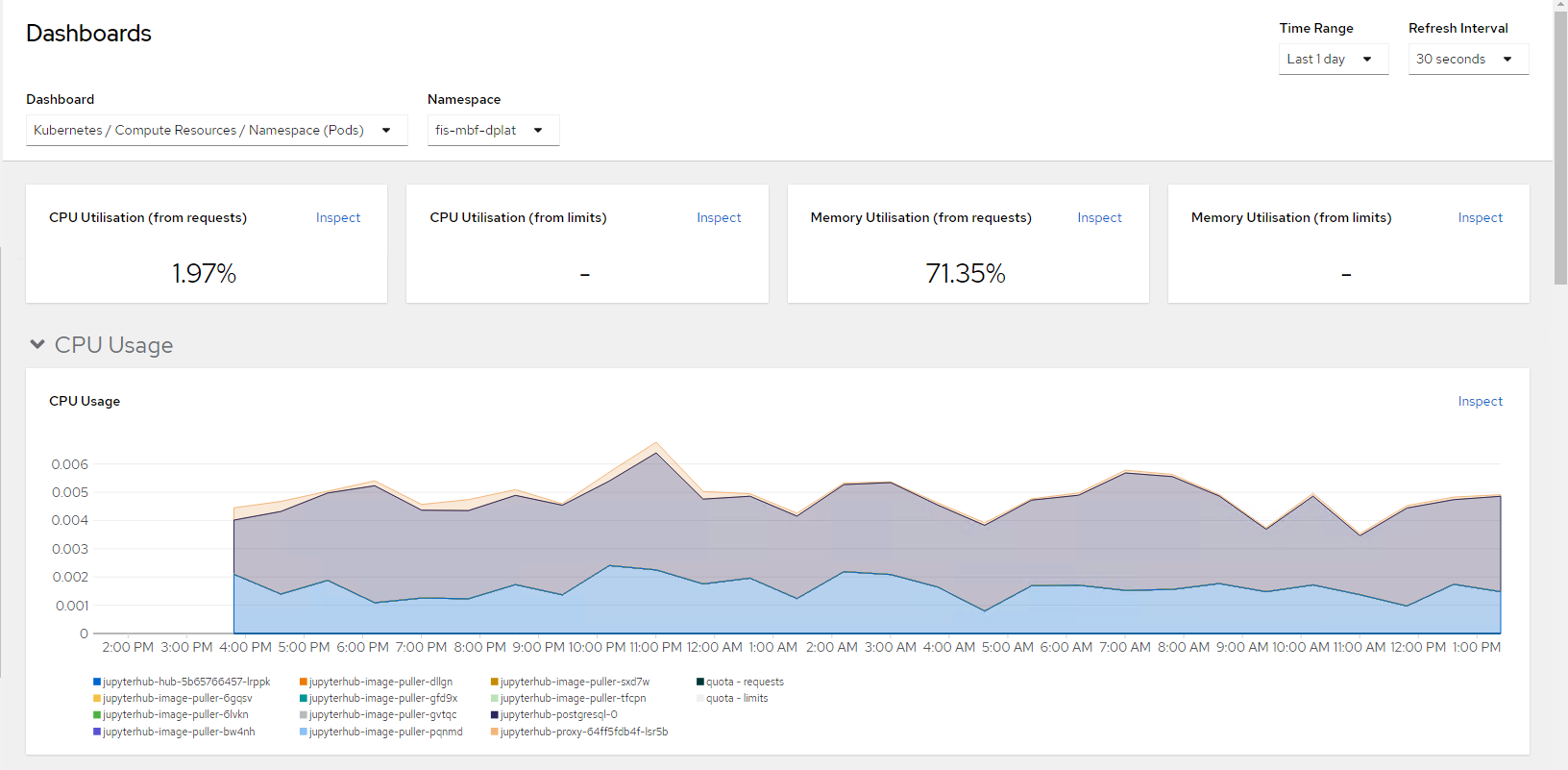
* + LDAP / Active Directory
  + OAuth: Google, Github, Azure Active Directory…
  + OpenID: Auth0, Keycloak

Đảm bảo tính an toàn bảo mật cho hệ thống, không thể xác thực bằng các tài khoản untrusted, anonymous

* 1. Monitoring

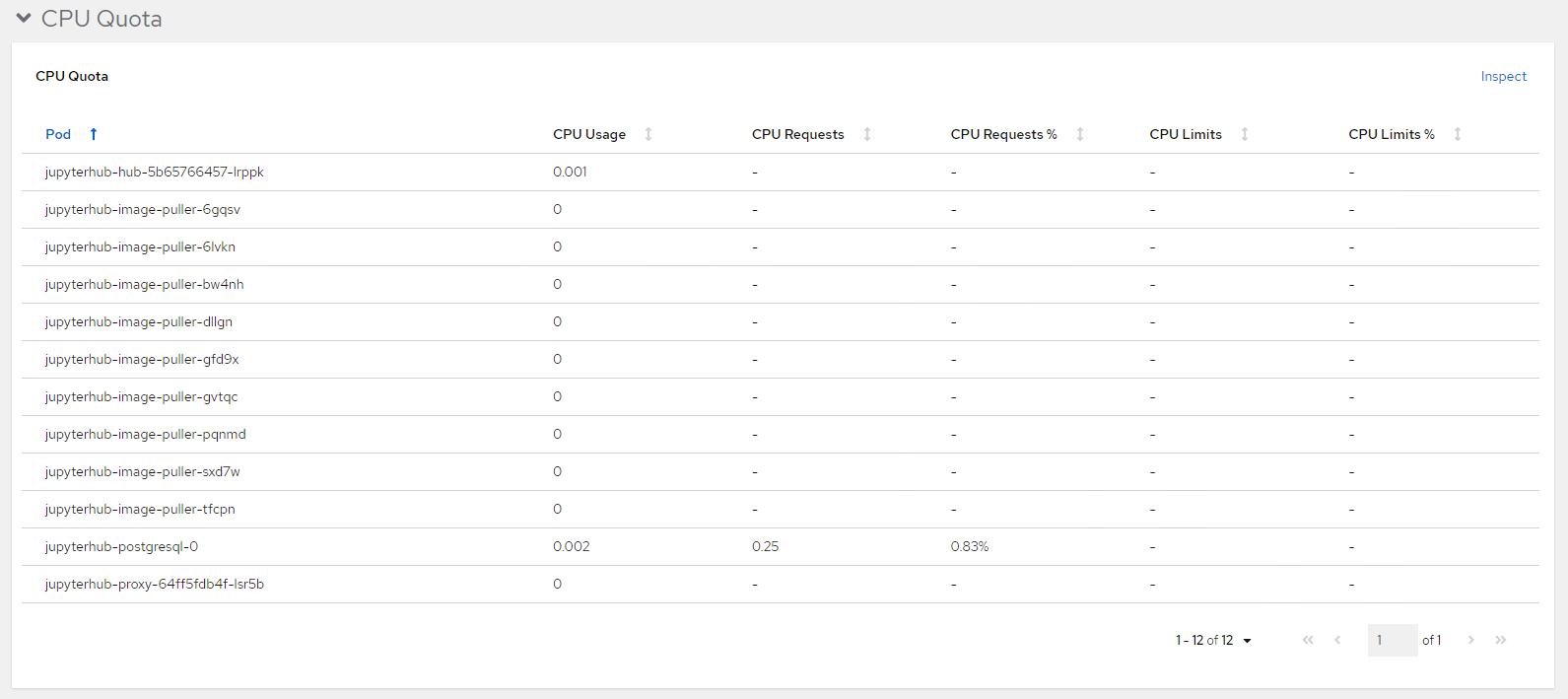
JupyterHub có thể monitoring qua giao diện của Openshift

* + Quản lí giám sát CPU Usage của Jupyter dưới dạng dashboard



Hình 13. Giám sát CPU Usage

* + Quản lí giám sát CPU Quota sử dụng (CPU đang sử dụng, CPU tối đa cho các service)



Hình 14. Giám sát CPU Quota

* + Quản lí giám sát Memory Usage của service và user pod

Graphical user interface, application, table

Description automatically generated

Hình 15. Giám sát Memory Usage

* + Quản lí giám sát Memory Quota sử dụng của servie và user pod (Memory đang sử dụng, Memory tối đa…)

A screenshot of a computer

Description automatically generated

Hình 16. Giám sát Memory Quota

* 1. Scalability

Hệ thống Jupyter được xây dựng trên nền tảng Openshift, chạy dưới hình thức containerize, vì vậy có khả năng Vertical Scaling và Horizontal Scaling

Vertical Scaling: Add thêm các node vào hệ thống để tăng tài nguyên sử dụng cho các pod

Horizontal Scaling: Tăng số lượng các pod chạy cùng một service trên Openshift, chia tải xử lý cho hệ thống

* 1. Availability

Hệ thống được xây dựng trên hạ tầng của Openshift, đảm bảo được tính sẵn sàng cao, các dịch vụ chạy 24/24 mà không gặp gián đoạn:

* + Cơ chế tạo Pod: Khi các pod dịch vụ chạy trên Openshift gặp lỗi, hoặc bị crash, hệ thống sẽ phát hiện và tạo một pod mới tương tự để đảm bảo hệ thống không bị gián đoạn
  + Persistent Volume (PV): Dữ liệu người dùng và các cấu hình hệ thống được lưu trữ tại các PV của hệ thống Openshift, các PV này được lưu trữ “persistent” vì vậy khi các pod bị xoá và tạo mới, dữ liệu người dùng vẫn nằm tại các PV và không thay đổi
  1. Backup and Restore

Vùng dữ liệu người dùng và dữ liệu cấu hình của Jupyter nằm trong các PV được lưu trên phân NFS share của hệ thống SAN storage

Quy trình backup và restore các dữ liệu này sẽ phụ thuộc vào việc sao lưu hệ thống SAN của Sacombank

* 1. Archiving

|  |  |  |
| --- | --- | --- |
| Data Retention |  |  |

Vùng dữ liệu NFS cho các Persistent Volume của Jupyter không có chính sách Retention, hiện tại đang để mặc định là không xoá hoặc dọn dẹp dữ liệu theo chu kì.

1. PHỤ LỤC

**Phụ lục danh sách các hình ảnh (mô hình) có trong tài liệu**

[Hình 1. Kiến trúc giải pháp 6](#_Toc119576716)

[Hình 2. Kiến trúc triển khai Dev/Test 7](#_Toc119576717)

[Hình 3. Kiến trúc triển khai UAT 10](#_Toc119576718)

[Hình 4. Kiến trúc triển khai Production 12](#_Toc119576719)

[Hình 5. Mô hình kiến trúc vật lý 15](#_Toc119576720)

[Hình 6. Mô hình kiến trúc luận lý 18](#_Toc119576721)

[Hình 7. Log service JupyterHub 21](#_Toc119576722)

[Hình 8. Log user JupyterHub 21](#_Toc119576723)

[Hình 9. Log service HubProxy 22](#_Toc119576724)

[Hình 10. Lỗi 503 Service Unavailable 22](#_Toc119576725)

[Hình 11. Lỗi thiếu tài nguyên 23](#_Toc119576726)

[Hình 12. Cấu hình HTTPS 443 24](#_Toc119576727)

[Hình 13. Giám sát CPU Usage 25](#_Toc119576728)

[Hình 14. Giám sát CPU Quota 25](#_Toc119576729)

[Hình 15. Giám sát Memory Usage 26](#_Toc119576730)

[Hình 16. Giám sát Memory Quota 26](#_Toc119576731)