

PROYECTO FINAL DE BIG DATA

Curso:	BIG DATA DEVELOPER
Profesor:	RUBEN QUISPE LLACCTARIMAY
Tema:	Tecnología para el Big Data
Estudiante:	Allan Mora Víquez
Fecha:	07/08/2023

Resultado de aprendizaje:

Elabora una solución de la arquitectura de construcción de un datalake corporativo, evidenciando la diferencia de las tecnologías de big data.

Diseña una arquitectura lógica funcional de Big Data desarrollando una solución de construcción de un datalake corporativo.

Instrucciones:

1. Lea atentamente el tema propuesto y el instrumento de evaluación, a fin de conocer los criterios de evaluación con los que será evaluado.
2. Analice el tema que se propone en el presente trabajo.
3. Realice una comparación de los 4 player AWS, AZURE, GCP y ORACLE.

Complete el **stack tecnológico de la arquitectura de Big data**, es decir, completar qué tecnología se debe utilizar en cada capa, tomado en cuenta los **principales players en el mercado de las plataformas cloud**: AZURE, AWS, GCP y ORACLE; asimismo, comparar qué nube es más confiable, a nivel precio, velocidad de carga, robustez, etc.; tomando en cuenta:

Comparación en el mismo momento temporal

- La más nueva suele vencer
- Cuidado con las comparativas «viejas» **No es sólo la funcionalidad:**
- Seguridad, robustez, carga y concurrencia, tiempo de respuesta • Precio
- Soporte (servicios vendor y ecosistema)
- Términos y condiciones

La clave es el caso de uso

- Es lo que va a priorizar la importancia de cada característica (incluyendo el precio)

Comparativas en Internet

- Busca y analiza con espíritu crítico

Exposición Caso I. Datalake Corporativo

Situación actual

Actualmente la organización no dispone de un sistema unificado de análisis con una visión integral de sus datos, orientada a la aportación de valor de negocio.

Existen soluciones basadas en tecnologías tradicionales. Consistentes en bases de datos relacionales (Oracle), procesos ETL de carga en scripts (PL/SQL) y explotación BI orientada un análisis retrospectivo de la información.

La organización no dispone de una solución de gobierno del dato corporativa, y los distintos sistemas funcionan como silos independientes, generando duplicidades, datos discrepantes y una gran dificultad para explotar conjuntamente la información procedente de los diversos sistemas.

A pesar de almacenar una gran cantidad de datos, los sistemas actuales no cubren las necesidades de la compañía ya que no cubren sus necesidades analíticas ni permiten la monetización de sus datos

Objetivos

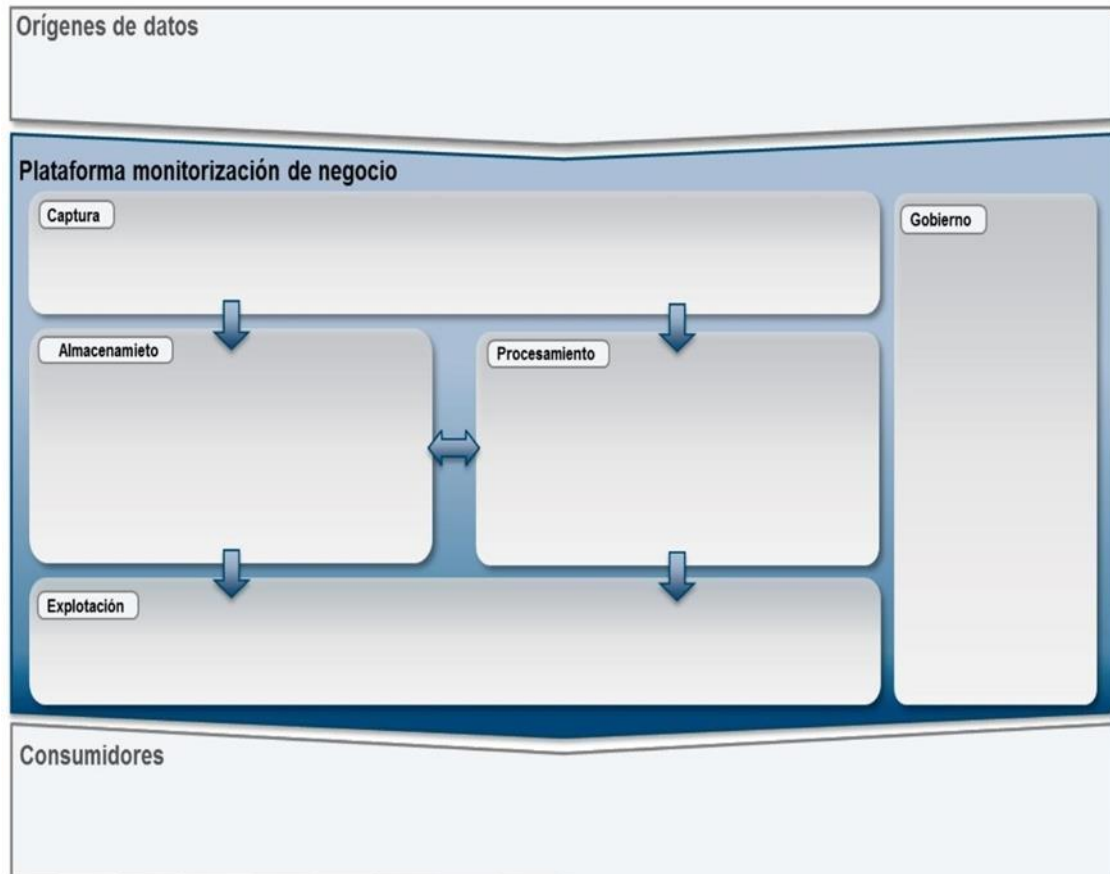
- Implantar un sistema unificado de gestión de la información para generar un impacto positivo sobre los diferentes procesos de negocio de la organización.
- Proporcionar una visión integrada de los indicadores de negocio más representativos para facilitar la monitorización de los procesos y la toma de decisiones.
- Disponer de los automatismos o mecanismos manuales de actuación para operar sobre los sistemas que permitan resolver incidencias o situaciones anómalas.
- Generar conocimiento de negocio a partir de análisis de la información histórica que mediante procesos de modelado y correlación permitan reconocer patrones y anticiparse de un modo proactivo a las futuras necesidades.

Enfoque

- ✚ Definición de un modelo flexible y escalable, basado en una arquitectura modular, que permita priorizar la incorporación de capacidades analíticas a la plataforma según las necesidades futuras e identificar los productos principales del stack tecnológico que darán respuesta a las necesidades funcionales de cada una de sus capas.
- ✚ Diseño de una arquitectura lógica de la información del Datalake que garantice la correcta disposición de la información en la plataforma según su naturaleza, facilite la seguridad y gestión de los accesos, etc.
- ✚ Definición de una solución corporativa de gobierno del dato que deberá ir acompañada de cambios en la estructura organizativa de la compañía e implicará la definición de nuevos roles y responsabilidades.
- ✚ Definición de una estrategia de implantación de la nueva plataforma y migración de los datos desde los sistemas analíticos actuales que minimice el impacto en la organización.

El proyecto es completar toda la solución de caso I, responder a los 4 enfoques, con la arquitectura respectiva, que distribución de hadoop eliges, completar el stack tecnológico por cada capa de la arquitectura.

Adjunto el cuadro para que completen el stack tecnológico respectivo por capa de la arquitectura.



Arquitectura lógica

■ ¿Cuántas áreas deberá tener la arquitectura lógica?

R/ En ese caso se diseño con 7 capas lógicas.

■ ¿Qué tipología de datos almacenará cada área? ¿Cuáles serán sus características?

R/ Las topologías y sus características serian:

- **Hadoop:** Para la capa de almacenamiento **raw** es una excelente opción ya que consigue procesar y almacenar grandes cantidades de datos utilizando hardware interconectado de bajo coste. Cientos o incluso miles de servidores dedicados de bajo coste trabajan juntos para almacenar y procesar datos dentro de un único ecosistema.
- **Google Cloud Platform (GCP):** Para la capa de almacenamiento **prepared data** de la infraestructura, proporciona servicio de almacenamiento de objetos de alta durabilidad que se puede escalar hasta exabytes de datos y servicios gestionados. Proporciona un acceso instantáneo a los datos desde cualquier servicio.
- Se utiliza **BigQuery** en la capa de **exploración** ya que es un almacén de datos de Google de bajo coste y totalmente administrado que permite extraer analíticas de petabytes de datos. Es autónomo, por lo que no es necesario gestionar ninguna infraestructura ni contar con un administrador de bases de datos.

■ ¿Quién serán los usuarios o procesos consumidores de cada área?

R/

- El área de Business Intelligence, a través de Data Studio
- Reportadores de autoservicio de las diferentes áreas a través de Looker Studio.
- Clientes atendidos a través de Chat-Boots
- Clientes desde aplicaciones móviles personalizadas.
- Web y Apps internas / externas personalizadas.
- Científicos de datos a través de Prediction Api, Cloud ML, Tense Flow y otros.
- Clientes internos haciendo Networking a través del ecosistema de Google.
- Usuarios de logística, marketing, Supply chain, interesados en el Internet de las cosas(IOT).

■ ¿Qué procesos o transformaciones se realizarán entre áreas?

R/ Una vez que las diferentes áreas de la empresa han migrado a Google Cloud Platform (GCP), es importante realizar procesos y transformaciones para aprovechar al máximo las capacidades de la nube y lograr una mayor eficiencia y colaboración en toda la organización y serian:

1. **Integración y Comunicación:** Facilitar la comunicación y colaboración entre las diferentes áreas de la empresa es esencial. Se debe implementar herramientas de comunicación y colaboración en la nube, como Google Workspace, para compartir documentos, colaborar en tiempo real y programar reuniones, también se pueden utilizar terceros como confluence para documentación y otros.
2. **Centralización de Datos:** La migración a la nube puede permitir la centralización de datos en un data lake. Esto permitirá a diferentes áreas acceder a datos relevantes de manera más eficiente, lo que puede ser útil para la toma de decisiones basadas en datos en toda la empresa.
3. **Automatización de Procesos:** Las áreas pueden beneficiarse de la automatización de procesos. Mediante la implementación de flujos de trabajo y procesos automatizados, es posible aumentar la eficiencia operativa y reducir errores.
4. **Análisis y Business Intelligence:** Aprovechar las capacidades analíticas de la nube para obtener información valiosa de los datos. Las áreas pueden colaborar en la generación de informes, análisis de tendencias y pronósticos para tomar decisiones más informadas.
5. **Desarrollo de Aplicaciones:** la migración a la nube puede impulsar una mayor colaboración entre los equipos de desarrollo y operaciones (DevOps). Puedes implementar prácticas de desarrollo ágil y utilizar servicios de desarrollo en la nube para acelerar el ciclo de desarrollo de aplicaciones.
6. **Seguridad y Cumplimiento:** La seguridad y el cumplimiento son preocupaciones importantes en cualquier migración a la nube. Diferentes áreas pueden colaborar para garantizar que las políticas y controles de seguridad estén implementados adecuadamente en toda la organización.
7. **Gestión de Recursos y Costos:** Es importante supervisar y gestionar los recursos en la nube de manera eficiente. Las áreas pueden colaborar para establecer políticas de asignación de recursos, controlar costos y optimizar el uso de servicios en la nube.
8. **Capacitación, Formación y acompañamiento:** Proporcionar capacitación y formación a los empleados de diferentes áreas sobre las herramientas y servicios en la nube utilizados puede ayudar a mejorar la adopción y el uso efectivo de la tecnología minimizando el impacto de la resistencia al cambio.
9. **Innovación y Experimentación:** La migración a la nube puede fomentar la innovación y la experimentación. Las áreas pueden colaborar en la identificación de nuevas oportunidades y en la implementación de proyectos piloto para probar nuevas ideas.

Se define una arquitectura lógica para el Datalake con tres áreas claramente diferenciadas:

■ Staging Data Area:

R/ Se agrego en el diseño en el área de almacenamiento la sección RAW para este fin.

■ Clean Data Area:

R/ Se agrego en el diseño el área de procesamiento donde ocurriría la limpieza de los datos.

■ PreparedData Area:

R/ Se agrego en el diseño el área de almacenamiento la sección PREPARED DATA para este fin.

Data Governance

■ ¿Qué cambios deberá impulsar la organización?

R/ La gobernanza de datos es un enfoque integral para administrar, proteger y utilizar datos de manera efectiva y responsable, la implementación exitosa de la gobernanza de datos implica cambios profundos en la cultura, las políticas, los procesos y las tecnologías de una organización.

1. **Cambio Cultural:** La gobernanza de datos implica un cambio cultural en toda la organización. Debe promoverse una mentalidad que valore la calidad de los datos, la colaboración y la toma de decisiones basada en datos. Todos los miembros de la organización deben entender la importancia de los datos precisos y confiables.
2. **Responsabilidad y Propiedad de los Datos:** Se deben establecer roles y responsabilidades claros para la gestión de datos. Esto implica designar propietarios de datos que sean responsables de definir, mantener y asegurar la calidad de los datos en sus áreas respectivas.
3. **Políticas y Normas de Datos:** La organización debe desarrollar y establecer políticas y normas claras para la creación, captura, almacenamiento, uso y distribución de datos. Estas políticas deben estar alineadas con los objetivos comerciales y las regulaciones relevantes.
4. **Metadatos y Catalogación:** La implementación de un sistema robusto de catalogación de datos y metadatos es esencial. Esto ayuda a rastrear la procedencia, el significado y la calidad de los datos, lo que facilita su búsqueda y comprensión.
5. **Privacidad y Cumplimiento:** La gobernanza de datos debe incluir medidas sólidas para proteger la privacidad de los datos y garantizar el cumplimiento de las regulaciones de protección de datos, como el Reglamento General de Protección de Datos (GDPR) en Europa.
6. **Calidad de Datos:** Se deben establecer procesos para monitorear y mejorar la calidad de los datos. Esto incluye la identificación y corrección de datos incorrectos o incompletos.
7. **Acceso y Seguridad:** La gobernanza de datos debe abordar el acceso y la seguridad de los datos. Se deben definir permisos de acceso apropiados para garantizar que solo las personas autorizadas puedan acceder y modificar los datos.
8. **Capacitación y Concienciación:** Los empleados deben recibir capacitación sobre las políticas y procedimientos de gobernanza de datos. Esto ayuda a garantizar que todos comprendan cómo manejar los datos de manera adecuada y respetar las normas.
9. **Tecnología y Herramientas:** La nube de Google ofrece un catálogo de datos totalmente gestionado y escalable para centralizar los metadatos y facilitar la búsqueda de datos. El catálogo de datos de Google se adherirá a los mismos controles de acceso que el usuario tiene sobre los datos (por lo que los usuarios no podrán buscar datos a los que no puedan acceder).
10. **Monitoreo y Auditoría:** La organización debe implementar mecanismos de monitoreo y auditoría para garantizar el cumplimiento continuo de las políticas de gobernanza de datos y la mejora continua de los procesos. Los registros de auditoría de Google Cloud pueden activarse para garantizar el cumplimiento de las auditorías en Google Cloud y responder a la pregunta "quién hizo qué, dónde y cuándo en todos los servicios de Google Cloud". Cloud Logging recopila automáticamente datos de los servicios de Google Cloud y usted puede alimentar los registros de las aplicaciones mediante el agente de Cloud Logging, FluentD o la API de Cloud Logging. Los registros de Cloud logging pueden reenviarse a GCS para archivarlos, a bigquery para analizarlos y también transmitirse a Pub/Sub para compartir registros con sistemas externos de terceros.

■ ¿Qué nuevos perfiles profesionales requerirá Power ON?

R/ La implementación exitosa de la gobernanza de datos en una organización puede requerir la incorporación de nuevos perfiles profesionales con habilidades y conocimientos específicos para liderar y gestionar los aspectos relacionados con los datos.

1. **Chief Data Officer (CDO):** El CDO es responsable de liderar la estrategia de gobernanza de datos en toda la organización. Supervisa la gestión, calidad, seguridad y privacidad de los datos, y trabaja para asegurar que los datos se utilicen de manera efectiva para impulsar la toma de decisiones y la innovación.
2. **Data Steward:** Los data stewards son responsables de la gestión y supervisión de datos específicos. Colaboran con los propietarios de datos y garantizan que los datos estén disponibles, sean precisos, estén actualizados y cumplan con las normas de calidad y cumplimiento.
3. **Data Quality Manager:** Este profesional se enfoca en garantizar la calidad de los datos. Supervisa la evaluación, monitoreo y mejora continua de la calidad de los datos, identificando problemas y coordinando acciones correctivas.

4. **Data Privacy Officer (DPO):** El DPO se encarga de garantizar que la organización cumpla con las regulaciones de privacidad de datos, como el GDPR. Supervisa las políticas y procedimientos de privacidad, coordina auditorías y maneja las solicitudes de privacidad de los datos.
5. **Data Governance Analyst:** Los analistas de gobernanza de datos trabajan en la implementación y ejecución de políticas y procesos de gobernanza de datos. Ayudan en la catalogación de datos, la documentación de metadatos y la monitorización de la calidad de los datos.
6. **Data Architect:** Los arquitectos de datos diseñan la estructura y la arquitectura de los sistemas de gestión de datos para garantizar que los datos sean accesibles, seguros y coherentes en toda la organización.
7. **Data Compliance Manager:** Este rol se centra en asegurar que la organización cumpla con las regulaciones y normas relacionadas con los datos, como las leyes de privacidad y protección de datos. Se encarga de mantener la conformidad en todas las operaciones de datos.
8. **Data Analyst:** Los analistas de datos trabajan con los datos para extraer información valiosa y proporcionar insights que apoyen la toma de decisiones. Su papel es fundamental para aprovechar al máximo los datos gestionados bajo la gobernanza.
9. **Change Management Specialist:** La implementación de la gobernanza de datos también puede requerir profesionales especializados en la gestión del cambio. Ayudan a la organización a adaptarse a las nuevas políticas y procesos de gobernanza de datos, asegurando una transición suave.
10. **Data Trainer / Educator:** Estos profesionales se encargan de proporcionar formación y educación a los empleados sobre las políticas, normas y prácticas de gobernanza de datos.

■ ¿Qué premisas deberá tener el diseño del nuevo modelo operativo?

R/ El diseño del nuevo modelo operativo para la gobernanza de datos debe basarse en una serie de premisas y considerar una amplia gama de actividades para garantizar una implementación efectiva y exitosa.

1. **Enfoque Estratégico:** La gobernanza de datos debe estar alineada con los objetivos estratégicos de la organización y respaldar su visión y misión.
2. **Colaboración Interdepartamental:** El modelo operativo debe promover la colaboración y la comunicación efectiva entre diferentes departamentos y equipos que gestionan y utilizan datos.
3. **Responsabilidad Claramente Definida:** Los roles y responsabilidades de la gobernanza de datos deben estar claramente definidos, y los propietarios de datos deben ser designados para asegurar una gestión adecuada.
4. **Cumplimiento Normativo:** El modelo debe garantizar el cumplimiento de las regulaciones y normas de privacidad y seguridad de datos aplicables.
5. **Enfoque en la Calidad de Datos:** La calidad de los datos debe ser una prioridad, y se deben establecer procesos para evaluar, monitorear y mejorar la calidad de los datos.
6. **Privacidad y Seguridad de Datos:** El modelo debe incluir medidas sólidas para proteger la privacidad y seguridad de los datos sensibles.
7. **Cambio Cultural:** La adopción exitosa de la gobernanza de datos requiere un cambio cultural que promueva la importancia de los datos y la colaboración en su gestión.
8. **Automatización y Tecnología:** Se debe aprovechar la automatización y las herramientas tecnológicas adecuadas para facilitar la implementación y el seguimiento de la gobernanza de datos.

■ ¿Qué actividades deberá contemplar?

R/

1. **Definición de Roles y Responsabilidades:** Identificar y definir los roles clave en el proceso de gobernanza de datos, como Data Stewards, Chief Data Officer, Data Privacy Officer, entre otros.
2. **Desarrollo de Políticas y Normas:** Crear políticas y normas claras que aborden la creación, captura, almacenamiento, uso y distribución de datos, así como la seguridad y la privacidad.
3. **Diseño de Procesos de Flujo de Datos:** Establecer procesos para la captura, transformación, almacenamiento y distribución de datos de manera coherente y eficiente.
4. **Implementación de Metadatos y Catalogación:** Definir y aplicar un sistema de catalogación de datos y metadatos para rastrear la procedencia y el significado de los datos.
5. **Evaluación y Mejora de la Calidad de Datos:** Diseñar procesos para evaluar y mejorar la calidad de los datos, incluida la identificación y corrección de problemas.
6. **Desarrollo de Herramientas y Tecnología:** Identificar y adoptar herramientas tecnológicas que respalden la gobernanza de datos, como sistemas de gestión de metadatos y soluciones de calidad de datos.

7. **Implementación de Formación y Capacitación:** Proporcionar formación y capacitación a los empleados sobre las políticas y procedimientos de gobernanza de datos.
8. **Establecimiento de Métricas y KPIs:** Definir métricas y Key Performance Indicators (KPIs) para medir el éxito de la implementación de la gobernanza de datos.
9. **Creación de Mecanismos de Cumplimiento:** Desarrollar procesos para garantizar el cumplimiento de las regulaciones de privacidad y seguridad de datos.
10. **Gestión del Cambio:** Diseñar estrategias para gestionar el cambio cultural y asegurar la adopción exitosa de la gobernanza de datos en toda la organización.

Estrategia de implantación

■ ¿Qué estrategia de migración seguiremos? ¿Es más conveniente un BigBang o en enfoque progresivo?

R/ La estrategia a seguir debe tener los siguientes pasos: Delimita el alcance del proyecto, Evalúa tus recursos, Diseño de la migración por áreas y etapas, Diseño de plan de pruebas prueba, Ejecución, Validación, Liberación. Además de estar basada en una planificación meticulosa debe hacerse con un conocimiento del negocio profundo, para evaluar todas las aristas y no dejar ninguna por fuera.

R/ Se seguirá el enfoque progresivo que busca es identificar todas las áreas del negocio, priorizarlas y planificarlas en la línea de tiempo y en etapas, también analizar cuales se pueden migrar de forma paralela sin que ocasionen cuellos de botella.

■ ¿Qué tareas deberemos realizar? ¿En qué orden? ¿Qué dependencias existen entre las tareas? ¿Se pueden paralelizar?

R/

Este sería el orden de los pasos para el plan de implementación:

1. **Evaluación de Requisitos:**
 - Identificar los datos que se migrarán y su estructura.
 - Definir los objetivos y beneficios esperados de la migración.
 - Establecer criterios de selección de la plataforma de nube.
2. **Diseño de la Arquitectura:**
 - Diseñar la arquitectura del Datalake, considerando factores como almacenamiento, procesamiento y seguridad.
 - Seleccionar las herramientas y servicios de la nube que se utilizarán.
3. **Planificación de la Migración:**
 - Establecer un cronograma y plazos para la migración.
 - Identificar los recursos humanos y técnicos necesarios.
 - Definir estrategias para la mitigación de riesgos.
4. **Seguridad y Cumplimiento:**
 - Diseñar políticas de seguridad y acceso a los datos en el Datalake.
 - Garantizar el cumplimiento de regulaciones y normativas aplicables.
5. **Diseño de Procesos ETL (Extract, Transform, Load):**
 - Planificar la extracción y transformación de datos desde las fuentes originales al Datalake.
 - Definir flujos de trabajo para cargar los datos en el Datalake de manera eficiente.

Pasos del Diseño del Plan por Área:

1. **Equipo y Roles:**
 - Asignar responsabilidades a los miembros del equipo.
 - Definir roles como arquitecto de datos, ingeniero ETL, analistas, etc.
2. **Infraestructura Tecnológica:**
 - Seleccionar la plataforma en la nube (por ejemplo, AWS, Azure, Google Cloud).
 - Diseñar la arquitectura de red y configuración de servidores.
3. **Gestión de Datos:**
 - Identificar las fuentes de datos y sistemas actuales.

- Diseñar el modelo de datos para el Datalake.
- 4. **Procesamiento y Análisis:**
 - Seleccionar herramientas para el procesamiento y análisis de datos en el Datalake.
 - Diseñar flujos de trabajo para el análisis de datos.
- 5. **Seguridad y Cumplimiento:**
 - Establecer políticas de acceso y control de datos.
 - Implementar medidas de encriptación y auditoría.

Pasos de la Ruta de Implementación:

1. **Configuración del Entorno de Nube:**
 - Crear y configurar la infraestructura en la nube.
 - Establecer conexiones seguras entre la infraestructura local y el Datalake en la nube.
2. **Desarrollo ETL:**
 - Diseñar y desarrollar los flujos de extracción, transformación y carga de datos.
 - Probar y optimizar los procesos ETL.
3. **Carga Inicial de Datos:**
 - Realizar la carga inicial de datos desde las fuentes existentes al Datalake.
4. **Desarrollo de Procesos de Análisis:**
 - Implementar flujos de trabajo para análisis y procesamiento de datos en el Datalake.
5. **Pruebas de Integración:**
 - Realizar pruebas exhaustivas de integración para asegurarse de que los datos se estén procesando y almacenando correctamente.

Pasos de la Ejecución:

1. **Migración Progresiva:**
 - Comenzar la migración de datos en etapas, priorizando conjuntos de datos críticos o de alto valor.
2. **Monitoreo y Optimización:**
 - Supervisar constantemente el rendimiento y la disponibilidad del Datalake.
 - Realizar ajustes según sea necesario para optimizar el rendimiento.
3. **Capacitación y Adopción:**
 - Capacitar al personal en el uso de las nuevas herramientas y flujos de trabajo.
 - Fomentar la adopción de la nueva plataforma.

Pasos de la Validación:

1. **Pruebas de Calidad de Datos:**
 - Verificar la integridad y precisión de los datos migrados.
 - Identificar y abordar problemas de calidad.
2. **Pruebas de Rendimiento:**
 - Evaluar el rendimiento de consultas y análisis en el Datalake.
 - Optimizar la arquitectura según los resultados.
3. **Validación de Seguridad:**
 - Revisar las políticas de seguridad y acceso para asegurarse de que cumplen con los estándares establecidos.
4. **Validación de Cumplimiento:**
 - Confirmar que la migración cumple con las regulaciones y normativas relevantes.

Pasos de la Validación por el Negocio por Área:

1. **Revisión de Requisitos Cumplidos:**
 - Colaborar con los representantes de cada área de negocio para verificar que los requisitos específicos de datos y procesos se hayan cumplido en la migración.
 - Obtener confirmación de que los datos necesarios para las operaciones comerciales están disponibles y son precisos en el nuevo entorno del Datalake.
2. **Pruebas de Uso de Datos:**
 - Realizar pruebas con los usuarios finales de cada área para asegurarse de que puedan acceder y utilizar los datos en el Datalake de manera efectiva.

- Validar que los informes, paneles y análisis que dependen de los datos migrados funcionen correctamente.
- 3. **Validación de Procesos de Negocio:**
 - Verificar que los flujos de trabajo y procesos de negocio que utilizan los datos migrados funcionen según lo esperado en el nuevo entorno.
 - Obtener comentarios de los usuarios sobre cualquier problema o mejora potencial.

Pasos del Visto Bueno de la Auditoría:

1. **Revisión de Seguridad y Cumplimiento:**
 - Proporcionar a la auditoría acceso a los sistemas y datos relevantes en el Datalake para realizar una revisión exhaustiva de las medidas de seguridad implementadas.
 - Compartir documentación detallada sobre políticas de acceso, encriptación y control de datos.
2. **Auditoría de Procesos y Flujos de Trabajo:**
 - Permitir que la auditoría evalúe los flujos de trabajo, procesos ETL y análisis de datos para verificar su precisión y coherencia.
 - Proporcionar explicaciones detalladas sobre cómo se realizan las transformaciones y cálculos en los datos.
3. **Validación de Cumplimiento Normativo:**
 - Presentar pruebas de que la migración y el uso de datos en el Datalake cumplen con las regulaciones y estándares aplicables a la industria y la empresa.
 - Responder a las preguntas y solicitudes de información de la auditoría.
4. **Informe de Auditoría y Aprobación:**
 - Revisar el informe de la auditoría con las recomendaciones y hallazgos.
 - Tomar medidas para abordar cualquier problema identificado durante la auditoría.
 - Obtener el visto bueno formal y la aprobación de la auditoría para proceder con la migración y el uso continuo del Datalake.

■ ¿Qué esfuerzo requerirá cada tarea? ¿Cuánto tiempo nos necesitaremos para llevarlas a cabo?
¿Con qué equipo?

R/ Todo proceso de migración de datos se puede realizar por etapas. Estas fases no tienen por qué presentar una configuración homogénea ya que muchas de ellas requieren más tiempo que otras para su ejecución, algunas presentan un nivel de dificultad mucho más elevado que el resto e incluso, en ocasiones, se considerará necesario repetir el desarrollo de una etapa hasta alcanzar la perfección que permita poder continuar adelante con garantías.

Tomando en cuenta lo anterior el tiempo estimado pueden variar entre 1 año 1 mes a 2 años y 2 mes, los plazos y esfuerzos son aproximados y pueden variar según las circunstancias específicas de la organización, esto puede extender el tiempo a 2.5 años.

Además, es de suma importancia contar con un equipo multidisciplinario con experiencia en migraciones a la nube, análisis de datos, gobierno de datos y desarrollo de soluciones para garantizar el éxito del proyecto.

1. **Evaluación y Planificación:**
 - Esfuerzo: Alto
 - Tiempo: 1-2 meses
 - Equipo: Arquitectos de soluciones, analistas de datos, líderes de proyecto
 - Descripción: En esta etapa, se debe evaluar el panorama actual de sistemas, datos y procesos. Se elaborará un plan detallado para la migración a GCP, identificando los sistemas a migrar, los requisitos de gobierno de datos y las necesidades analíticas.
2. **Diseño de la Arquitectura en la Nube:**
 - Esfuerzo: Alto
 - Tiempo: 1-2 meses
 - Equipo: Arquitectos de soluciones, ingenieros en la nube

- Descripción: Diseñar la arquitectura en la nube que cumpla con los requisitos de la compañía y permita la integración de datos de diferentes sistemas. Se definirán las instancias, bases de datos, herramientas de análisis y flujos de datos.
- 3. **Migración de Datos y Sistemas:**
 - Esfuerzo: Alto
 - Tiempo: 3-6 meses (dependiendo de la complejidad)
 - Equipo: Ingenieros de datos, administradores de bases de datos, ingenieros en la nube
 - Descripción: Migrar los datos y sistemas existentes a GCP. Esto puede incluir la reestructuración de bases de datos, la transformación de datos y la integración de sistemas. Se deben asegurar la consistencia y calidad de los datos.
- 4. **Implementación de Gobierno de Datos:**
 - Esfuerzo: Medio a Alto
 - Tiempo: 2-4 meses
 - Equipo: Analistas de datos, especialistas en gobierno de datos
 - Descripción: Establecer un marco de gobierno de datos en GCP. Definir políticas de calidad de datos, metadatos, roles y permisos, y procesos para asegurar la integridad y coherencia de los datos en la nube.
- 5. **Desarrollo de Análisis y Monetización:**
 - Esfuerzo: Medio a Alto
 - Tiempo: 3-6 meses
 - Equipo: Analistas de datos, científicos de datos, desarrolladores
 - Descripción: Desarrollar soluciones de análisis y visualización de datos en la nube. Esto puede incluir la creación de cuadros de mando, informes y herramientas de análisis para permitir la toma de decisiones basada en datos y la monetización de los mismos.
- 6. **Capacitación y Soporte:**
 - Esfuerzo: Bajo a Medio
 - Tiempo: 3-6 meses
 - Equipo: Equipo de soporte, especialistas en formación
 - Descripción: Capacitar al personal en el uso de las nuevas soluciones en la nube y proporcionar soporte continuo para resolver problemas y optimizar el uso de la plataforma.

■ ¿Cómo gestionaremos el cambio en la organización?

Estrategia de Migración

1. ¿Qué proveedor eligen AZURE, AWS, GCP?

R/ Se eligió **GCP** por ser un ecosistema con muchas herramientas de punta, ser muy estables y de bajo costo (El más bajo de los tres).

2. ¿Cuál es la estrategia de migración a GCP?

R/

La estrategia de la migración será dividirla en etapas y áreas de negocio las cuales se priorizarán para su pase a la nube con validaciones y firma de aceptación por parte del negocio una vez que se valla entregando cada área hasta completar la etapa correspondiente.

3. ¿Qué servicios de Azure, GCP, AWS, Oracle utilizará en el pipeline de Big Data?

Diseño de la arquitectura lógica para la migración de Oracle a la plataforma Google Cloud Platform (GCP) junto con el stack tecnológico de Big Data para cada una de las capas lógicas:

Infraestructura: Plataforma Big Data en GCP:

- Hadoop: Para infraestructura de los datos raw es una excelente opción ya que consigue procesar y almacenar grandes cantidades de datos utilizando hardware interconectado de bajo coste. Cientos o incluso

miles de servidores dedicados de bajo coste trabajan juntos para almacenar y procesar datos dentro de un único ecosistema.

- Google Cloud Platform (GCP): La base de la infraestructura, proporciona recursos escalables y servicios gestionados.

Fuentes: Orígenes de Datos:

- Bases de Datos Relacionales: Oracle Database es un sistema de gestión de bases de datos objeto-relacional, se conoce comúnmente como Oracle RDBMS o simplemente Oracle.
- La base de la plataforma Oracle Business Intelligence Suite Enterprise Edition es un servidor de BI diseñado para ser altamente escalable.

Captura: Recolección e Ingesta de Datos:

- Apache Airflow que es una excelente opción, esta es probablemente la herramienta más utilizada para orquestar y programar sus canales de datos más utilizada.
- Se utilizará con Python y PySpark para hacer las lecturas validaciones e ingestas de datos.

Almacenamiento: Persistencia de la Información:

- Hadoop es una estructura de software de código abierto para almacenar datos y ejecutar aplicaciones en clústeres de hardware comercial. Proporciona almacenamiento masivo para cualquier tipo de datos, enorme poder de procesamiento y la capacidad de procesar tareas o trabajos concurrentes virtualmente ilimitados.
- Google Cloud Storage: Almacenamiento escalable y duradero para datos en bruto y procesados.
- Google BigQuery: Data warehouse y motor de consulta para análisis interactivo de grandes conjuntos de datos.

Procesado: Transformación de Datos:

- Google Cloud Dataproc: Servicio de Apache Spark y Apache Hadoop administrado para procesamiento de datos.
- Google Cloud Dataflow: Con Apache Beam para procesamiento de datos en lotes y en tiempo real.

Explotación: Explotación de la Información:

- Google BigQuery: Para consultas ad-hoc y análisis interactivo.
- Google Cloud Datalab: Para explorar, analizar, transformar y visualizar datos y construir modelos de aprendizaje automático.
- Google Data Studio: Para visualización y creación de paneles.
- Google Looker: Para análisis y generación de informes.

Gobierno: Gobierno de la Información y Control:

- Google Cloud Identity and Access Management (IAM): Control de acceso a recursos.
- Google Cloud Audit Logging: Registro de auditoría para actividades en la nube.
- Google Cloud Data Catalog: Catalogación y descubrimiento de datos.
- Google Cloud DLP: Protección de datos y prevención de pérdida de información.

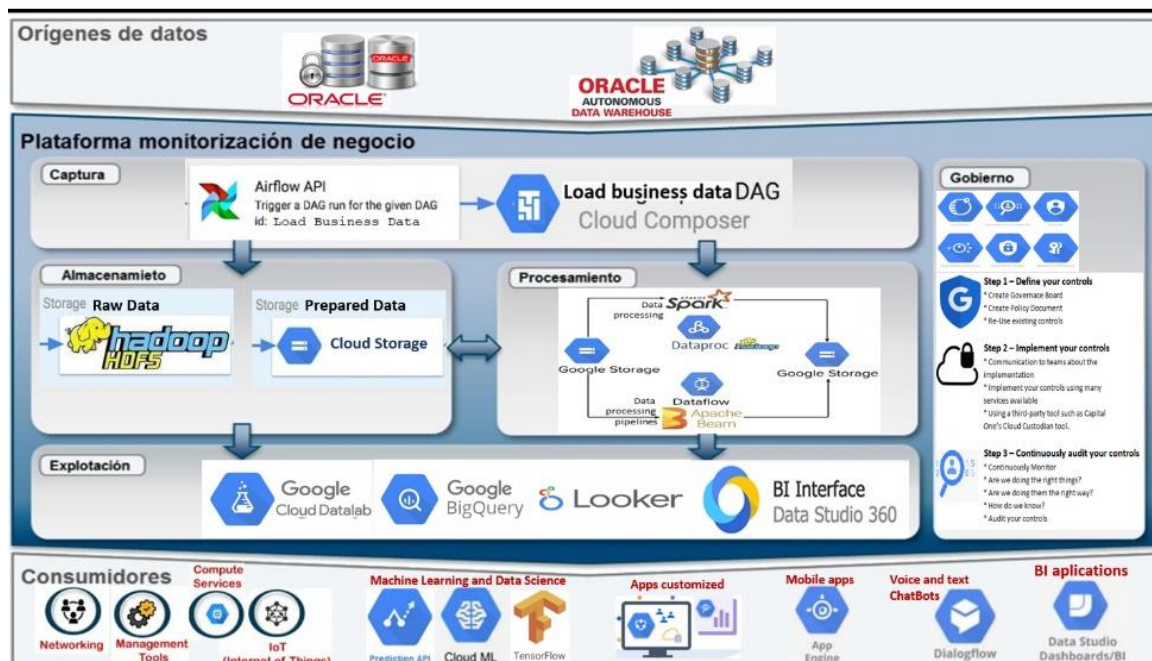
Google Cloud Platform ofrece una amplia gama de servicios que se pueden combinar de diversas formas para crear una arquitectura de Big Data escalable y eficiente.

La presente arquitectura se diseñó teniendo en cuenta los flujos de datos, el rendimiento, la seguridad y los requisitos de cumplimiento.

4. Diseñar la arquitectura de la solución con los servicios utilizados.

Resolución Caso I. Datalake Corporativo

Stack tecnológico



5. ¿Cómo se verificar o validan los datos de On Premise y Cloud?
R/ Se deben de considerar los siguientes aspectos:

Validación de Datos On-Premise:

1. **Análisis Exploratorio de Datos (EDA):** Se debe realizar un análisis detallado de los datos en tu entorno on-premise. Identifica patrones, valores atípicos, valores nulos y otras anomalías que puedan afectar la calidad de los datos.
2. **Pruebas Unitarias y de Integración:** Crear pruebas automatizadas para verificar la precisión y consistencia de los datos en diferentes sistemas y bases de datos on-premise. Esto puede incluir pruebas de integridad referencial, verificación de cálculos y validación de formatos.
3. **Comparación de Datos:** Comparar los datos en diferentes sistemas para detectar discrepancias y duplicidades. Utiliza herramientas de comparación de datos para identificar diferencias entre conjuntos de datos y sistemas.

Validación de Datos en la Nube (Cloud):

1. **Migración y Carga de Datos:** Durante la migración a la nube, se debe verificar que los datos se hayan transferido correctamente y que no se hayan producido pérdidas ni alteraciones. Realiza pruebas de carga y migración para garantizar que los datos se encuentren en la nube según lo previsto.
2. **Pruebas de Calidad de Datos en la Nube:** Utiliza herramientas y scripts automatizados para ejecutar pruebas de calidad de datos en la nube. Verifica la coherencia, la integridad y la precisión de los datos en la nueva plataforma.
3. **Validación de Procesos de Transformación:** Con los ETL, se debe asegurar que se estén aplicando correctamente. Realiza pruebas para verificar que los datos se estén transformando según lo planeado y que los resultados sean consistentes.
4. **Monitoreo Continuo:** se deben implementar sistemas de monitoreo continuo para detectar y alertar sobre problemas de datos en la nube. Esto puede incluir la supervisión de cambios en los datos, la detección de anomalías y la generación de alertas cuando se identifiquen problemas.

5. **Validación de Resultados de Análisis:** Si se está realizando análisis de datos en la nube, se debe asegurar de validar los resultados obtenidos. Compara los resultados con los datos originales y verifica que los cálculos y las conclusiones sean precisos y coherentes.

6. ¿Estimar un precio del diseño e implementación de un data lake corporativo?

R/

Estimar el precio del diseño e implementación de un Data Lake corporativo puede ser complicado debido a la variedad de factores que influyen en los costos. Los precios pueden variar según la ubicación geográfica, la escala del proyecto, las tecnologías utilizadas, el nivel de personalización y otros factores.

Pero haciendo una aproximación general de los elementos que debes considerar al estimar el precio:

Recursos Humanos: El costo del personal necesario para diseñar, desarrollar e implementar el Data Lake, incluyendo arquitectos de datos, ingenieros de datos, desarrolladores, analistas de datos, projects managers, product owners, expertos en calidad de datos, business analysts, scrum masters y especialistas en gobernanza de datos.

1. **Tecnologías y Herramientas:** Los costos de las licencias de software, suscripciones a servicios en la nube, herramientas de análisis, bases de datos y otras tecnologías necesarias para construir y operar el Data Lake.
2. **Infraestructura:** Los costos asociados con la infraestructura, ya sea en la nube o en entornos locales, como servidores, almacenamiento, redes y otros componentes de hardware.
3. **Consultoría Externa:** Si la organización requiere asesoramiento externo o servicios de consultoría para el diseño y la implementación del Data Lake, estos también deben tenerse en cuenta en el presupuesto.
4. **Capacitación y Soporte:** Los costos de capacitar a los empleados para usar y administrar el Data Lake, así como el soporte continuo para resolver problemas y optimizar el sistema.
5. **Gestión de Cambios y Comunicación:** Los costos asociados con la gestión del cambio en la organización y la comunicación interna sobre el nuevo Data Lake.
6. **Escalabilidad y Mantenimiento:** Se debe considerar los costos a largo plazo para mantener, escalar y mejorar el Data Lake a medida que cambian las necesidades de la organización y se incorporan nuevas fuentes de datos.

7. Estimar el tiempo de diseño e implementación del Datalake.

R/ Como se respondió en la sección de “Estrategia de implantación”

Tomando en cuenta lo anterior el tiempo estimado pueden variar entre 10 meses a 1 año y 6 meses, los plazos y esfuerzos son aproximados y pueden variar según las circunstancias específicas de tu organización, esto puede extender el tiempo a 2 años.

8. Puntos a agregar o desarrollar

Respondiendo a la parte 3: Realice una comparación de los 4 player AWS, AZURE, GCP y ORACLE.

R/






Estos cuatro gigantes de la computación en nube tienen sus pros y sus contras.

Dado que los objetivos y requisitos de cada empresa son diferentes de los de las demás, también lo es el proveedor de servicios en la nube que elige. La selección del proveedor de servicios en la nube para un particular depende de su público específico y de sus objetivos empresariales. Los tres AWS, Azure y GCP ofrecen precios, planes, capacidades de almacenamiento y servicios de recuperación de datos competitivos.

Si su empresa maneja grandes cargas de trabajo e información confidencial, AWS es la opción perfecta para usted.

Sin embargo, si no está tan seguro de los servicios en la nube y está pensando en iniciarse en la nube pública, entonces Azure es la mejor alternativa de nube híbrida.

Por otra parte, si estás manejando un negocio nativo de la nube y quieres gestionar todo sin romper el banco, entonces debes priorizar Google Cloud Platform.

Soluciones, servicios y herramientas	-Datalake on AWS -AWS Lake Formations -Amazon S3	-HDInsight -Data Lake Analytics -Azure Data Lake Storage	-Data Lake Modernization -BiLake, DataFlow, BigQuery,	-Data motion and integration -Data lake -Data lakehouse
Virtual Machines	AWS EC2	Azure Virtual Machines	Google Compute Engine	Oracle VM VirtualBox
PAAS	Elastic Beanstalk	Cloud Services	App Engine	Cloud Services 21
Managed Kubernetes services	AKS	EKS	GKE	OKE
IAM para un acceso controlado	Comparing AWS	Azure	Google Cloud IAM services	Access Management (IAM)
Serverless	AWS Lambda	Azure Functions	Google Cloud Functions	OCI functions
Object Storage	S3	Block Blob	Cloud Storage	(OCI) Object Storage
Archive Storage	Glacier	Archive Storage	Coldline	(OCI) Object Storage
FileStorge	EFS	Azure files	ZFS/Avere	(OCI) File Storage
zona de disponibilidad múltiple sla	99,99%	99,99%	99,99%	99,99%
Zona de disponibilidad de servidor único sla	90% por Hora	99% (Premium ssd) 99.5%(SSD)	99,5%	99,5%
Data Warehouse	Redshift	Sql Warehouse	BigQuery	Warehouse Management (WMS)
NoSQL databases	Cosmos DB	DynamoDB	Cloud Datastore and Bigtable	-Oracle Real Application Clusters (RAC) -Oracle Autonomous Database -Oracle Exadata Cloud Service

Cuadro comparativo.