**1. Planificación y Diseño**

* **Definir Requisitos**: Clarifica los objetivos y los casos de uso específicos.
* **Elegir Herramientas y Tecnologías**: Selecciona las herramientas basadas en los requisitos.
* **Diseño de la Arquitectura**: Crea un diagrama detallado de la arquitectura.

**2. Infraestructura**

* **Hardware**: Decide entre servidores locales y la nube.
* **Servicios en la Nube**: AWS, Google Cloud, Azure.
* **Contenedores**: Docker para empaquetar aplicaciones.
* **Orquestación de Contenedores**: Kubernetes para gestionar la escalabilidad y el despliegue de contenedores.

**3. Ingesta de Datos**

* **Apache NiFi**: Para ingesta y enrutamiento de datos en tiempo real.
* **Apache Sqoop**: Para transferir datos entre bases de datos relacionales y Hadoop.
* **Apache Kafka**: Sistema de mensajería para ingesta en tiempo real.

**4. Almacenamiento**

* **HDFS**: Sistema de archivos distribuido para almacenamiento masivo.
* **Apache Cassandra**: Base de datos NoSQL para almacenamiento de datos en tiempo real.
* **Amazon S3**: Almacenamiento en la nube.

**5. Procesamiento de Datos**

* **Apache Spark**: Procesamiento en tiempo real y por lotes.
* **Apache Flink**: Procesamiento de datos en tiempo real.
* **Apache Beam**: Modelo unificado para procesamiento por lotes y en tiempo real.
* **PySpark**: Procesamiento de datos usando Python en Spark.

**6. Gobernanza y Seguridad**

* **Apache Atlas**: Gestión de metadatos y linaje de datos.
* **Apache Ranger**: Gestión de políticas de acceso y seguridad.
* **Cifrado**: Datos en reposo y en tránsito.

**7. Análisis y Machine Learning**

* **Apache Mahout**: Algoritmos de aprendizaje automático.
* **TensorFlow**: Modelos avanzados de aprendizaje profundo.
* **MLlib**: Librería de aprendizaje automático de Spark.

**8. Visualización y BI**

* **Grafana**: Monitorización y visualización en tiempo real.
* **Metabase**: BI y dashboards interactivos.
* **Apache Superset**: Visualización y creación de dashboards.

**9. Orquestación de Tareas**

* **Apache Airflow**: Orquestación y planificación de flujos de trabajo.

**10. Monitorización y Logging**

* **ELK Stack (Elasticsearch, Logstash, Kibana)**: Ingesta, búsqueda y visualización de logs.
* **Prometheus**: Monitorización y alertas.

**11. Automatización y DevOps**

* **CI/CD Pipelines**: Jenkins para automatización de despliegues.
* **Infrastructure as Code (IaC)**: Terraform o Ansible para gestión de infraestructura.

**12. Despliegue y Configuración**

* **Configuración de Entornos**: Desarrollo, prueba, producción.
* **Gestión de Configuraciones**: Uso de herramientas como Ansible o Chef.
* **Pruebas**: Pruebas unitarias, de integración y de rendimiento.

**13. Documentación y Capacitación**

* **Documentación Técnica**: Guías de usuario, diagramas, especificaciones técnicas.
* **Capacitación**: Formación para el equipo en las herramientas y tecnologías utilizadas.

**Pasos Detallados para el Despliegue**

1. **Configurar Infraestructura en la Nube**
   * Provisión de máquinas virtuales o contenedores.
   * Configuración de redes y seguridad.
2. **Instalar y Configurar Hadoop HDFS**
   * Configurar NameNode y DataNodes.
   * Configurar HDFS-site.xml y Core-site.xml.
3. **Instalar y Configurar Apache Kafka**
   * Configurar brokers y zookeeper.
4. **Instalar y Configurar Apache NiFi**
   * Configurar flujos de datos y procesadores.
5. **Instalar y Configurar Apache Spark**
   * Configurar clúster de Spark.
   * Desplegar aplicaciones PySpark.
6. **Instalar y Configurar Apache Airflow**
   * Configurar la base de datos backend.
   * Configurar DAGs para flujos de trabajo.
7. **Configurar Almacenamiento Adicional (Cassandra, S3)**
   * Instalar y configurar nodos de Cassandra.
   * Configurar S3 buckets.
8. **Implementar Herramientas de Gobernanza y Seguridad**
   * Configurar Apache Atlas para metadatos.
   * Configurar Apache Ranger para seguridad.
9. **Configurar Herramientas de Visualización**
   * Instalar Grafana y configurar dashboards.
   * Instalar Metabase y conectar con las fuentes de datos.
10. **Configurar Monitorización y Logging**
    * Configurar ELK stack.
    * Configurar Prometheus para monitorización.
11. **Desarrollar y Desplegar Modelos de Machine Learning**
    * Desarrollar modelos en TensorFlow o MLlib.
    * Desplegar modelos para inferencia en tiempo real o por lotes.
12. **Automatizar Despliegues y Gestión de Configuraciones**
    * Configurar pipelines de CI/CD con Jenkins.
    * Gestionar infraestructura con Terraform o Ansible.