Escuela
Politécnica
Superior

Análisis de Métricas de Películas y Series mediante la construcción de un almacén de datos con tecnología Big Data

Grado en Ingeniería Informática

Trabajo de fin de grado

Autor:
Alejandro Seguí Apellániz

Tutor:
Alejandro Reina Reina

Julio 2025

Universitat d'Alacant
Universidad de Alicante

Resumen

El gran volumen y la variedad de contenidos audiovisuales nos genera un desafío: ¿cómo analizar información útil de millones de películas y series para entender sus tendencias y factores de éxito? Existen páginas que tienen información de contenidos audiovisuales, pero no permiten realizar ningún tipo de análisis con una cantidad masiva de datos audiovisuales.

Este trabajo aborda esa problemática diseñando e implementando una arquitectura Big Data capaz de integrar y analizar a gran escala información procedente de diferentes fuentes.

Para ello se recopilan y se enriquecen datasets con información procedente de IMDb y JustWatch, dichos datasets contiene millones de datos sobre diferentes variables de los contenidos audiovisuales y dichos contenidos abarcan desde 1974 hasta 2031, es decir que hay contenidos que aún no se han estrenado, como por ejemplo Avatar 5 que se estrena en diciembre de 2031, (aunque los análisis se realizarán hasta 2024).

Para este proyecto se utilizan diferentes herramientas Big Data como Apache Spark para el procesamiento ETL en Scala, Hive como base de datos, y Power BI para la visualización.

Tras el análisis realizado, podemos observar que las producciones nacionales se han reducido en los últimos años y que, por el contrario, las plataformas de streaming como Netflix o Prime Video han aumentado exponencialmente su catálogo de contenidos.

Motivación, justificación y objetivo general

La principal motivación de llevar a cabo este proyecto es el interés que me surgió en el área del Big Data tras haber cursado la asignatura de INGP en donde utilizaban este tipo de herramientas para realizar análisis. También abarcan temas como el Business Intelligence.

Elegir el contenido cinematográfico como tema principal en este proyecto, se debe a diferentes razones. En primer lugar, las películas llevan siendo una fuente de entretenimiento, cultura y arte desde su invención, con un impacto en diferentes países y generaciones. Además, el cine es un área que se encuentra en constante evolución donde la mezcla de tecnología y creatividad permite innovaciones en la producción y distribución del contenido audiovisual.

Cabe destacar que el análisis de películas no solo se refiere al ámbito cinematográfico, sino que también abarca diferentes campos. Por ejemplo, puede tener un enfoque histórico, las películas muestran acontecimientos, cambios sociales de la época. La industria cinematográfica es un mercado a nivel mundial que genera miles de millones de euros.

El objetivo principal que tiene el TFG es mediante el uso de herramientas de Big Data poder analizar información de diferentes aspectos sobre contenidos audiovisuales. Además, se utilizarán técnicas de webscraping para recoger datos, permitiendo obtener más información relevante.

Agradecimientos

En primer lugar, quiero agradecer a mi tutor Alejandro Reina que me ha orientado y apoyado a lo largo de este proyecto, siempre que he tenido algún problema o duda a la hora de la realización del proyecto, Alejandro me ha contestado de manera rápida y precisa lo que facilitó el desarrollo y la finalización del proyecto.

También me gustaría agradecer a mi familia que han sido un pilar fundamental para mí en los momentos difíciles y han confiado en mí hasta el final. El apoyo que me han dado me ha ayudado a superar los retos que se me han presentado a lo largo de estos años en la universidad.

A continuación, me gustaría agradecer a los profesores que me han dado clase durante la carrera. Cada uno con su forma de explicar, ofreciendo sus conocimientos y su experiencia han ayudado en ampliar mis conocimientos.

También, me gustaría agradecer a mis compañeros que me han acompañado durante estos años. Su compañía y el apoyo mutuo han ayudado a que este camino sea más llevadero.

Este trabajo ha sido parcialmente financiado por los proyectos: AETHER-UA (PID2020-112540RB-C43) del Ministerio de Ciencia e Innovación de España, ValgrAI Project de la Escuela Valenciana de Postgrado en Inteligencia Artificial. The BALLADEER (PROMETEO/ 2021/088) Project de la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital. Alejandro Reina Reina cuenta con una Beca de Doctorado Industrial cofinanciada por la Universidad de Alicante y la Spin-off Lucentia Lab.

Citas

Los datos son el nuevo petróleo.

- Clive Humby

La ciencia de hoy es la tecnología del mañana.

- Edward Teller

La inteligencia es la capacidad de adaptarse al cambio.

- Stephen Hawking

Sin datos, solo eres otra persona con una opinión.

- W. Edwards Deming

La información es el oxígeno del mundo moderno.

- Ronald Reagan

Índice de contenidos

1. Introducción 15
2. Metodología 16
2.1 Planificación del proyecto 16
2.2 Copia de seguridad del proyecto 17
2.3 Control de versiones 18
2.4 Instantáneas Virtual Box 18
3. Objetivos 19
4. Estado del arte 23
4.1 Evolución histórica y relevancia del análisis audiovisual 23
4.2 Herramientas y plataformas de análisis audiovisual 24
4.3 Introducción al Business Intelligence 25
4.3.1 Definición Business Intelligence 25
4.3.2 Importancia y ventajas del BI 26
4.4 Arquitectura BI 26
4.4.1 Fuentes de datos 27
4.4.2 Procesos ETL 29
4.4.3 Almacén de datos 30
4.4.4 Servidor de consultas (OLAP) 32
4.4.5 Visualización o Dashboard 33
4.5 Problemas de la arquitectura BI 34
4.6 Historia del Big Data 35
4.7 Tecnologías Big Data 36
4.7.1 Landscape Big Data 37
4.7.2 Herramientas de procesamiento 37
4.7.2.2 Spark 37
4.7.2.2.1 Spark MLlib 38
4.7.2.2.2 Spark Streaming 39
4.7.2.2.3 Spark SQL 39
4.7.2.2.4 Spark GraphX 40
4.7.2.3 Databricks 40
4.7.3 Herramientas de streaming y mensajería 41
4.7.3.1 Apache Flink 41
4.7.3.2 Apache Kafka 43

4.7.4 Herramientas de almacenamiento y BD 44
4.7.4.1 Hive 44
4.7.4.2 Cassandra 46
4.7.4.3 MongoDB 48
4.7.4.4 HBase 49
4.7.4.5 Cassandra vs MongoDB vs HBase 50
4.7.5 Herramientas de gestión de clúster 51
4.7.5.1 Hadoop 51
4.7.5.1.1 HDFS 53
4.7.5.1.2 MapReduce 54
4.7.5.1.3 YARN 55
4.7.5.2 Apache Ambari 56
4.7.6 Herramientas ETL 57
4.7.6.1 Pentaho Data Integration 57
4.7.6.2 Pipelines de ETL 58
4.7.7 Herramientas de Cubos OLAP 58
4.7.7.1 Microsoft SQL Server Management Studio (SSMS) 58
4.7.7.2 Kylin 59
4.7.7.3 Kylin vs Microsoft SQL Server Management Studio 61
4.7.8 Herramientas de desarrollo 61
4.7.8.1 Zeppelin Notebook 61
4.7.8.2 Jupyter Notebook 62
4.7.8.3 Databricks Notebooks 63
4.7.8.4 Visual Studio 63
4.7.8.5 Notebooks vs IDEs 64
4.7.9 Virtualización y contenedores 65
4.7.9.1 Máquinas virtuales 65
4.7.9.2 Docker 65
4.7.10 Herramientas de visualización y BI 66
4.7.10.1 Power BI 66
4.7.10.2 Tableau 68
4.7.10.3 Qlick Sense 68
5. Diseño de la arquitectura 69
5.1 Esquema de la arquitectura 69

5.2 Herramientas utilizadas 69
6. Implementación del sistema 73
6.1 Fuentes de datos 73
6.1.1 IMDb – Non Commercial Datasets 73
6.1.1.1 title.basics.tsv 74
6.1.1.2 title.rating.tsv 74
6.1.1.3 name.basics.tsv 75
6.1.2 IMDb – webscraping 76
6.1.2.1 title.basics.tsv 76
6.1.2.2 name.basics.tsv 77
6.1.3 JustWatch 78
6.1.3.1 netflix.csv, prime.csv y appletv.csv 78
6.2 Diseño de almacén de datos 79
6.2.1 Diseño conceptual 79
6.2.2 Diseño lógico 82
6.3 Procesos ETL 84
6.3.1 Extracción de datos por webscraping 84
6.3.1.1 IMDb 85
6.3.1.2 JustWatch 86
6.3.2 Transformación y carga de datos 88
6.3.2.1 Dimensión tipo 88
6.3.2.2 Dimensión plataforma 88
6.3.2.3 Dimensión producción 90
6.3.2.4 Dimensión empresa productora 91
6.3.2.5 Dimensión género 92
6.3.2.6 Dimensión idioma 93
6.3.2.7 Dimensión premio más importante 94
6.3.2.8 Dimensión lanzamiento 95
6.3.2.9 Dimensión actor 96
6.3.2.10 Dimensión director 99
6.3.2.11 Dimensión guionista 100
6.3.2.12 Hecho contenido audiovisual 101
6.3.2.13 Carga tablas Hive 103
7. Cubos OLAP 107

7.1 Diseño cubo OLAP 107
7.2 Despliegue y procesamiento 110
8. Cuadros de mandos 113
8.1 Dashboard, objetivos y KPIs 114
8.1.1 Análisis de las producciones de cine nacional 117
8.1.1.1 KPI 1.1: Comprobar si el crecimiento de películas españolas aumenta en un +30% del 2021 al 2024. 119
8.1.1.2 KPI 1.2: Comprobar si el crecimiento de series españolas aumenta en un +30% del 2021 al 2024. 120
8.1.2 Análisis de las producciones más populares 122
8.1.2.1 KPI 2.1: Analizar si las películas populares no han bajado su producción en un -15% en los últimos 20 años [2004 – 2024] 124
8.1.2.2 KPI 2.2: Analizar si pertenecen al género "Drama" al menos un 10% de las series populares en los últimos 20 años [2004 – 2024] 126
8.1.3 Análisis de los lanzamientos en fechas clave de España y EE. UU 127
8.1.3.1 KPI 3.1: Analizar que los contenidos de España lanzados en "Navidad" aumentan +15% en los últimos 20 años [2004 – 2024] 129
8.1.3.2 KPI 3.2: Analizar que los contenidos de España lanzados en "Verano" aumentan +20% en los últimos 20 años [2004 – 2024] 129
8.1.3.3 KPI 3.3: Analizar que los contenidos de EE. UU lanzados en "Otoño" aumentan +20% en los últimos 20 años [2004 – 2024] 131
8.1.3.4 KPI 3.4: Analizar que los contenidos de EE. UU lanzados en "Primavera" aumentan +20% en los últimos 20 años [2004 – 2024] 131
8.1.4 Análisis de los contenidos de Netflix 134
8.1.4.1 KPI 4.1: Analizar que el 50% de los contenidos de Netflix sea contenido procedente de EE. UU de 2017 a 2019 135
8.1.4.2 KPI 4.2: Analizar que en total en Netflix se producen al menos 80 películas procedente de EE. UU de 2017 a 2019 135
8.1.4.3 KPI 4.3: Analizar que en total en Netflix se producen al menos 30 series animadas de 2017 a 2019 136
8.1.5 Análisis del reconocimiento en premios Oscar 137
8.1.5.1 KPI 5.1: Analizar si los contenidos premiados con Oscar aumentan en 5 unidades en los últimos 20 años [2004 – 2024] 139
8.1.5.2 KPI 5.2: Analizar si en total hay al menos 100 contenidos que han ganado algún premio Oscar en los últimos 20 años [2004 – 2024] 140
8.2 Preguntas analíticas 141
8.2.1 ¿Qué géneros dominan en la actualidad? 141

8.2.2 ¿Qué géneros tienen una mayor recaudación? 142
8.2.3 ¿Qué películas tienen una mayor recaudación y mayor puntuación? 143
8.2.4 ¿En qué épocas se realiza más contenidos actualmente? 144
8.2.5 ¿Hay un auge de producciones en nuevos países en los últimos 10 años? 145
8.2.6 ¿De qué países son los actores más populares (estrellas consagradas)? 146
8.2.7 ¿A qué grupo de edad pertenecen los actores conocidos? 147
8.2.8 ¿Un director ganador de algún Oscar influye en la puntuación de la película? 148
8.2.9 ¿Qué idiomas originales son los más frecuentes en los contenidos? 149
8.2.10 ¿Qué plataformas tienen los contenidos mejores valorados? 150
8.2.11 ¿Cómo ha crecido el volumen de contenidos de las plataformas de streaming en los últimos 20 años? 151
9. Conclusiones 153
10. Posibles mejoras futuras 155
10.1 Enriquecer la información con fuentes externas 155
10.2 Escalabilidad con un clúster distribuido 155
10.3 Inteligencia artificial 155
11. Opinión personal 156

Índice de figuras

Figura 1. Trello tareas del proyecto - Elaboración propia 16
Figura 2. Ejemplo de tarea creación ETLs - Elaboración propia 17
Figura 3. Backup disco duro externo - Elaboración propia 17
Figura 4. Repositorio GitHub. Sección ETLs - Elaboración propia 18
Figura 5. Instantáneas Virtual Box - Elaboración propia 18
Figura 6. Arquitectura BI - wordpress.com 27
Figura 7. Esquema estrella – dataprix.com 31
Figura 8. Ejemplo cubo OLAP - dba.stackexchange.com 33
Figura 9. Ejemplo Dashboard – evotic.es 34
Figura 10. 3V Big Data - researchgate.net 35
Figura 11. Big Data Landscape 2024 - pracdata.io 37
Figura 12. Librerías Spark - medium.com 38
Figura 13. Arquitectura Apache Flink - apache.org 42
Figura 14. Arquitectura Apache Kafka - interviewbit.com 44
Figura 15. Arquitectura Apache Hive. forum.huawei.com 45
Figura 16. Estructura Apache Hive - developer.aliyun.com 46
Figura 17. Ecosistema de Hadoop - apache.org 52
Figura 18. Ecosistema HDFS - medium.com 54
Figura 19. Ejemplo Map Reduce - codesyntax.com 55
Figura 20. Ecosistema YARN - hadoop.apache.org 56
Figura 21. Ejemplo flujo ETL - Elaboración propia 58
Figura 22. Estructura Apache Kylin - todobi.com 60
Figura 23. Ejemplo Dashboard Power BI - microsoft.com 67
Figura 24. Diagrama de arquitectura BI - blogvisionarios.com 69
Figura 25. Ejemplo carga dataframe usurios – Elaboración propia 71
Figura 26. Ejemplo select usuarios – Elaboración propia 71
Figura 27. Datos de nacimiento de un actor a extraer - imdb.com 77
Figura 28. Diseño conceptual - Elaboración propia 80
Figura 29. Diseño lógico - Elaboración propia 83
Figura 30. Tablas en la base de datos de Hive - Elaboración propia 84
Figura 31. Scrapeo información adicional profesionales - Elaboración propia 86
Figura 32. Scrolleo simulado por Selenium WebDriver de contenidos de Netflix - justwatch.com 87

Figura 33. Scrapeo información adicional contenidos en plataformas – Elaboración propia 87
Figura 34. Tabla tipo - Elaboración propia 88
Figura 35. Tabla plataforma - Elaboración propia 89
Figura 36. Tabla producción - Elaboración propia 91
Figura 37. Tabla empresa productora - Elaboración propia 92
Figura 38. Tabla género - Elaboración propia 93
Figura 39. Tabla idioma - Elaboración propia 94
Figura 40. Tabla premio más importante - Elaboración propia 95
Figura 41. Tabla lanzamiento - Elaboración propia 96
Figura 42. Histograma premios actores - Elaboración propia 97
Figura 43. Histograma grupo edad actores - Elaboración propia 98
Figura 44. Tabla actor - Elaboración propia 99
Figura 45. Tabla director - Elaboración propia 100
Figura 46. Tabla guionista - Elaboración propia 101
Figura 47. Tabla hechos contenidoAudiovisual (parte 1) - Elaboración propia 103
Figura 48. Tabla hechos contenidoAudiovisual (parte 2) - Elaboración propia 103
Figura 49. Creación conector ODBC – Elaboración propia 104
Figura 50. Creación database SQL – Elaboración propia 105
Figura 51. Carga de tablas del almacén de datos 105
Figura 52. Vistas tablas – Elaboración propia 106
Figura 53. Conexión BD – Elaboración propia 107
Figura 54. Carga vistas – Elaboración propia 108
Figura 55. Dimensiones creadas – Elaboración propia 108
Figura 56. Jerarquía temporal – Elaboración propia 109
Figura 57. Medidas y dimensiones cubo – Elaboración propia 109
Figura 58. Relaciones dimensiones y tabla de hechos – Elaboración propia 110
Figura 59. Carga datos cubo – Elaboración propia 111
Figura 60. Premios por género – Elaboración propia 112
Figura 61. Relaciones claves ajenas Power BI – Elaboración propia. 113
Figura 62. Página principal del dashboard – Elaboración propia 115
Figura 63. Visualización contenidos nacionales – Elaboración propia 117
Figura 64. Gráficos tipo y plataformas contenidos nacionales 2021 - 2024 – Elaboración propia 118
Figura 65. Gráfico géneros nacionales más producidos 2021-2024 – Elaboración propia 118
Figura 66. Gráfico evolución películas nacionales 2021-2024 – Elaboración propia 119

Figura 67. Evolución anual de producción de películas españolas – Elaboración propia. 119

Figura 68. Gráfico evolución series nacionales 2021-2024 – Elaboración propia 120

Figura 69. Evolución anual de producciones de españolas – Elaboración propia..... 121

Figura 70. Visualización contenidos populares – Elaboración propia. 122

Figura 71. Gráficos tipo y plataformas contenidos populares 2004 - 2024 – Elaboración propia 123

Figura 72. Gráfico géneros contenidos más populares 2004-2024 – Elaboración propia. 123

Figura 73. Gráfico evolución películas populares 2004-2024 – Elaboración propia 124

Figura 74. Evolución producciones contenidos populares – Elaboración propia..... 125

Figura 75. Gráfico evolución serie dramas populares y porcentaje 2004-2024 – Elaboración propia 126

Figura 76. Visualización lanzamiento de contenidos en España y EE. UU – Elaboración propia. 127

Figura 77. Gráfico tipo y cantidad contenidos España y EE. UU 2004 - 2024 – Elaboración propia 128

Figura 78. Gráfico géneros con más contenidos en fechas clave en España y EE. UU 2004-2024 – Elaboración propia..... 128

Figura 79. Gráfico lanzamiento contenidos en Navidad y Verano en España 2004-2024 – Elaboración propia..... 129

Figura 80. Evolución producciones de España en Navidad y Verano – Elaboración propia. 130

Figura 81. Gráfico lanzamiento contenidos en Otoño y Primavera en EE. UU 2004-2024 – Elaboración propia..... 131

Figura 82. Evolución producciones de EE. UU en Otoño y Primavera – Elaboración propia. 132

Figura 83. Visualización contenidos Netflix 2017-2019 – Elaboración propia. 134

Figura 84. Gráfico tipo y cantidad contenidos de EE. UU en Netflix 2017 - 2019 – Elaboración propia..... 135

Figura 85. Gráfico cantidad películas EE. UU de Netflix 2017-2019 – Elaboración propia 136

Figura 86. Gráfico cantidad series animadas de Netflix 2017-2019 – Elaboración propia137

Figura 87. Contenidos ganadores de Oscar 2004-2024 – Elaboración propia. 137

Figura 88. Cantidad contenidos premiados 2004 -2024 – Elaboración propia. 138

Figura 89. Géneros ganadores de Oscar 2004 – 2024 – Elaboración propia..... 138

Figura 90. Gráfico evolución producción contenidos con Oscar 2004-2024 – Elaboración propia 139

Figura 91. Suma contenidos totales con premios Oscar – Elaboración propia 140

Figura 92. Gráfico géneros más producidos – Elaboración propia 141

Figura 93. Gráfico géneros populares 2014-2024 – Elaboración propia..... 142
Figura 94. Géneros que producen más ingresos – Elaboración propia 142
Figura 95. Contenidos con mayor recaudación – Elaboración propia..... 143
Figura 96. Contenidos más populares en la audiencia – Elaboración propia 144
Figura 97. Gráfico épocas lanzamiento 2014-2024 – Elaboración propia 145
Figura 98. Gráfico países producción populares – Elaboración propia..... 145
Figura 99. Gráfico países producción populares 2014-2024 – Elaboración propia 146
Figura 100. Países actores consagrados – Elaboración propia 146
Figura 101. Grupo edad actores conocidos entre 2014 y 2024 – Elaboración propia..... 148
Figura 102. Puntuación media contenidos cuando un director tiene y no tiene Oscar – Elaboración propia..... 149
Figura 103. Idiomas originales más frecuentes contenidos – Elaboración propia 150
Figura 104. Gráfica puntuación por plataforma – Elaboración propia..... 151
Figura 105. Gráfica cantidad de contenidos por plataforma – Elaboración propia..... 152

Figura Anexo 1. Comprobación versión java..... 161
Figura Anexo 2. Descarga Spark..... 161
Figura Anexo 3. Descomprimir paquete Spark 162
Figura Anexo 4. Configuración de variables de entorno..... 162
Figura Anexo 5. Ejecución comando pyspark..... 163
Figura Anexo 6. Funcionamiento de Spark 163
Figura Anexo 7. Fichero .bashrc 164
Figura Anexo 8. Instalación ssh 165
Figura Anexo 9. Fichero core-site.xml..... 166
Figura Anexo 10. Fichero hdfs-site.xml..... 166
Figura Anexo 11. Fichero mapred-site.xml..... 166
Figura Anexo 12. Fichero yarn-site.xml..... 167
Figura Anexo 13. Generación clave 167
Figura Anexo 14. Hadoop activado..... 168
Figura Anexo 15. Instalación mysql-server..... 169
Figura Anexo 16. Inicio mysql-server..... 169
Figura Anexo 17. Privilegios mysql-server..... 170
Figura Anexo 18. Cambiar ubicación mysql 170
Figura Anexo 19. Fichero .bashrc 170
Figura Anexo 20. Fichero hive-env.sh 171

Figura Anexo 21. Fichero hive-site.xml 171
Figura Anexo 22. Hive en funcionamiento 171
Figura Anexo 23. Fichero zeppelin-site.xml 172
Figura Anexo 24. Cambio de puerto zeppelin-env 172
Figura Anexo 25. Zeppelin funcionando con Spark. 172
Figura Anexo 26. Proceso Zeppelin en Spark 173

Índice de tablas

Tabla 1. Metadatos title.basics.tsv 74
Tabla 2. Metadatos title.rating.tsv 75
Tabla 3. Metadatos name.basics.tsv 75
Tabla 4. Datos extraídos por webscraping sobre contenidos 77
Tabla 4. Datos extraídos por webscraping sobre profesionales 78
Tabla 6. Metadatos netflix.csv, prime.csv y appletv.csv 79
Tabla 7. Categorización de actores por premios ganados 97
Tabla 8. Categorización de actores por edad 98

1. Introducción

Podemos definir Big Data como la tecnología que se utiliza para analizar grandes volúmenes de datos. El conjunto de datos con el que vamos a trabajar es tan grande y complejo que las herramientas tradicionales de gestión de datos no pueden manejarlos de manera eficiente. Estos datos pueden venir de diversas fuentes, como datasets públicos, páginas webs, etc. Permitiendo obtener información que nos puede ser útil para llevar a cabo diferentes análisis.

En la actualidad, la capacidad de gestionar grandes volúmenes de datos ha revolucionado diversos sectores, como la industria cinematográfica que es el tema principal del TFG. El Big Data, no solo ha transformado el modus operandi de las empresas, sino también está influyendo en la toma de decisiones basadas en los datos recogidos.

En la industria cinematográfica, el Big Data tiene una gran importancia, ya que permite analizar diferentes parámetros como gustos de los espectadores, evaluar el impacto de las películas y optimizar su distribución tanto en cines como en plataformas de streaming, lo que permite llegar a un público muy amplio.

El problema que se presenta en este proyecto es gestionar la información relacionada con el contenido de contenido audiovisual que incluyen películas y series, con el uso de la tecnología Big Data. De este tema se puede obtener una cantidad de datos enorme y podríamos decir que casi infinita, ya que existen millones de reproducciones cinematográficas y prácticamente a diario se van creando nuevos contenidos audiovisuales. Por lo que el uso de tecnología Big Data es fundamental para realizar este análisis.

2. Metodología

Al enfrentarnos a un proyecto que tiene una mayor envergadura que otros trabajos que hemos tenido en otras asignaturas del grado, es importante tener una metodología de trabajo, la cual nos facilite la organización de nuestro trabajo. Para ello utilizaremos varias herramientas, que nos permitan realizar un control de nuestro trabajo.

2.1 Planificación del proyecto

Para llevar a cabo la planificación de nuestro proyecto, usaremos Trello, la cual es una herramienta online que nos permite organizar diferentes tareas de nuestro proyecto, además de poder marcarnos fechas límites para poder realizarlas. A continuación, podemos ver el repositorio Trello creado [1].

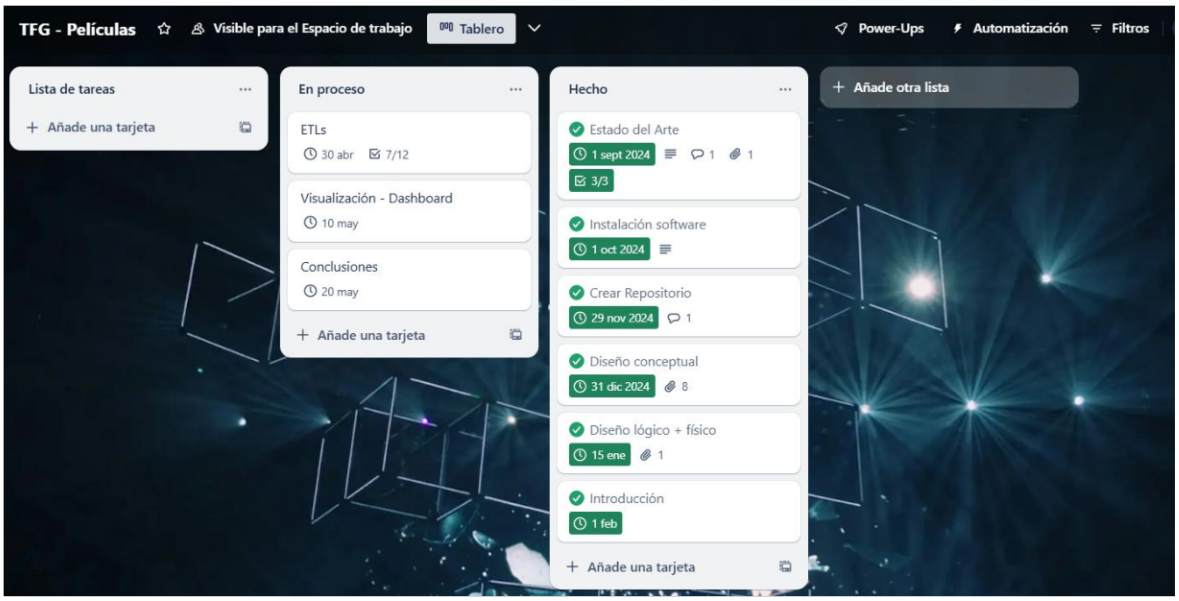


Figura 1. Trello tareas del proyecto - Elaboración propia

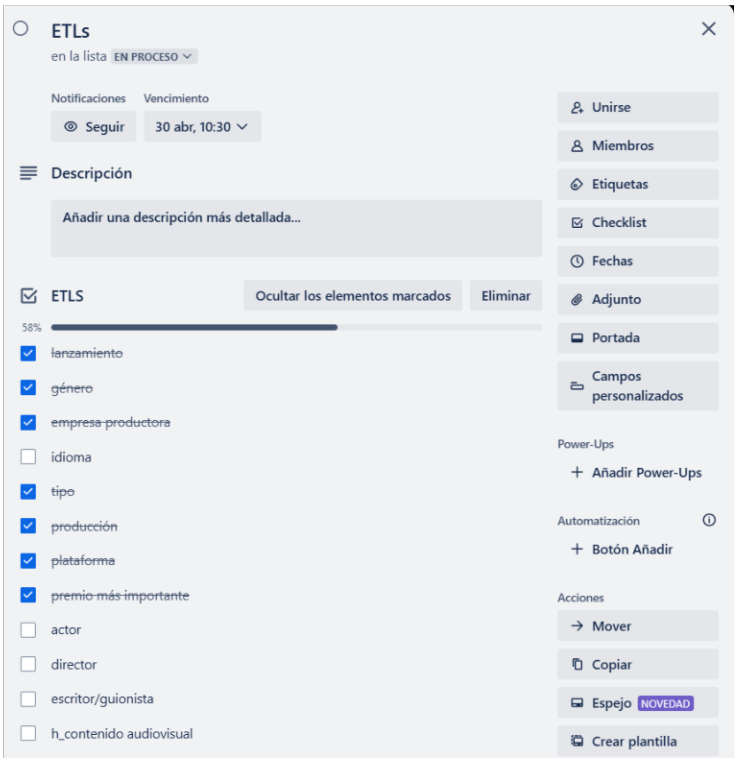


Figura 2. Ejemplo de tarea creación ETLs - Elaboración propia

2.2 Copia de seguridad del proyecto

Para asegurarme de no perder todo el trabajo que he ido haciendo, por algún problema que pueda surgir con mi ordenador, me he asegurado de hacer una copia de seguridad de todo lo que esté relacionado con el trabajo como la máquina virtual, archivos generados, o incluso la misma memoria. Para hacer esta copia de seguridad he utilizado un disco duro externo, en el cual como dije hice copia de seguridad de todo lo que esté relacionado con este proyecto. A continuación, podemos ver una imagen de la copia de seguridad.

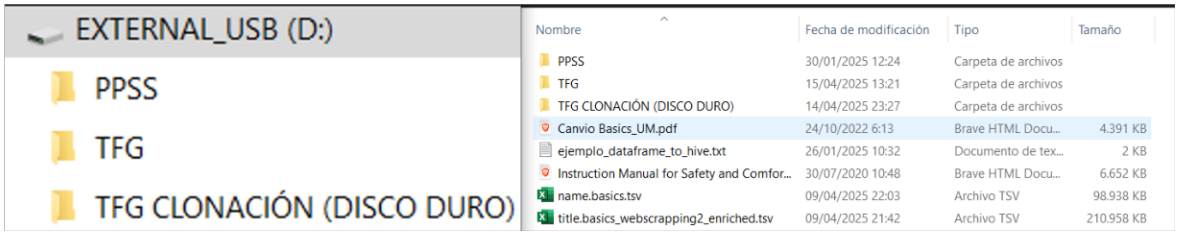


Figura 3. Backup disco duro externo - Elaboración propia

2.3 Control de versiones

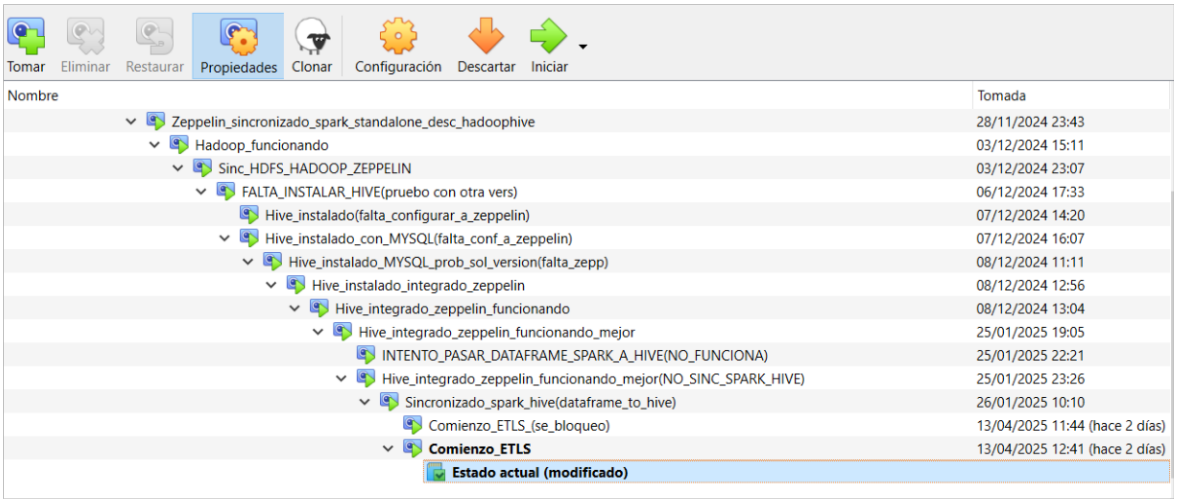
En cuanto a herramienta de control de versiones he usado GitHub. Esta herramienta nos será de gran utilidad para subir el código de nuestros ETLs, o cualquier otro fichero que no queramos perder. Además, si queremos continuar trabajando en otro computador en el que no tengamos estos ficheros simplemente descargaremos el repositorio remoto y ya podremos continuar con nuestro proyecto.

Name	Last commit message
..	
etl_empresaProductora.scala	Etls
etl_genero.scala	Etls
etl_lazamiento.scala	Etls
etl_plataforma.scala	Etls
etl_premiomasimportante.scala	Etls
etl_produccion.scala	Etls
etl_tipo.scala	Etls

Figura 4. Repositorio GitHub. Sección ETLs - Elaboración propia

2.4 Instantáneas Virtual Box

Para instalar el software y montar el almacén de datos utilizaremos la herramienta Virtual Box con el sistema operativo Ubuntu. Virtual Box nos permite crear instantáneas, están son copias de seguridad del sistema operativo hasta el punto en el cual hacemos dicha instantánea, lo cual es útil por si se produce algún error en nuestro sistema, nos permitiría restaurarlo hasta el punto en el que hicimos la instantánea.



Tomar | Eliminar | Restaurar | Propiedades | Clonar | Configuración | Descartar | Iniciar

Nombre	Tomada
Zeppelin_sincronizado_spark_standalone_desc_hadoophive	28/11/2024 23:43
Hadoop_funcionando	03/12/2024 15:11
Sinc_HDFS_HADOOP_ZEPPELIN	03/12/2024 23:07
FALTA_INSTALAR_HIVE(pruebo con otra vers)	06/12/2024 17:33
Hive_instalado(falta_configurar_a_zeppelin)	07/12/2024 14:20
Hive_instalado_con_MYSQL(falta_conf_a_zeppelin)	07/12/2024 16:07
Hive_instalado_MYSQL_prob_sol_version(falta_zepp)	08/12/2024 11:11
Hive_instalado_integrado_zeppelin	08/12/2024 12:56
Hive_integrado_zeppelin_funcionando	08/12/2024 13:04
Hive_integrado_zeppelin_funcionando_mejor	25/01/2025 19:05
INTENTO_PASAR_DATAFRAME_SPARK_A_HIVE(NO_FUNCIONA)	25/01/2025 22:21
Hive_integrado_zeppelin_funcionando_mejor(NO_SINC_SPARK_HIVE)	25/01/2025 23:26
Sincronizado_spark_hive(dataframe_to_hive)	26/01/2025 10:10
Comienzo_ETLS_(se_bloqueo)	13/04/2025 11:44 (hace 2 días)
Comienzo_ETLS	13/04/2025 12:41 (hace 2 días)
Estado actual (modificado)	

Figura 5. Instantáneas Virtual Box - Elaboración propia

3. Objetivos

El objetivo principal de este proyecto es analizar la evolución y los factores de éxito de diferentes contenidos audiovisuales, esto servirá de guía para posible toma de decisiones en futuras producciones.

Además, para que la producción de futuros contenidos se base en una evidencia sólida, se han planteado unos objetivos específicos que buscan analizar tanto la presencia nacional como el rendimiento y la visibilidad de los contenidos a nivel internacional. A continuación, planteamos cada uno de los objetivos, acompañados de sus KPIs SMART (específicos, medibles, alcanzables, relevantes y acotados en el tiempo), y al final se proponen una serie de preguntas analíticas las cuales el sistema debe ser capaz de dar respuesta y estas preguntas serán respondidas posteriormente junto al dashboard.

Cabe destacar que para el objetivo de analizar las producciones españolas se ha tomado como referencia el Plan Fomento del cine español 2021 – 2024 del ICAA (Instituto de la Cinematografía y de las Artes Audiovisuales), el cual se fija un aumento del 30% de los contenidos 2021 a 2024, con el fin de aumentar la presencia de contenidos nacionales después de la crisis del COVID [2].

Para el segundo objetivo se buscará analizar cuál es la tendencia y estabilidad de la producción de los contenidos más populares en los últimos 20 años. Los contenidos en IMDb tienen una puntuación de 0 a 10 por lo que analizaremos aquellos contenidos que tienen al menos una puntuación de 8. También es importante destacar que un contenido puede tener una puntuación de 10 y por ejemplo 5 votos, por los contenidos analizados también deben tener al menos 1000 votos.

A continuación, se pretende comprobar, si en los últimos 20 años ha habido un aumento sostenido en los estrenos en fechas clave en España como "Navidad" y "Verano". Dado que las fechas de estreno que se tienen con los datos que vamos a trabajar son las fechas de estreno en España, definiremos "Verano" como los meses junio, julio y agosto y "Navidad" como diciembre. No obstante, lo habitual en EE. UU es concentrar los lanzamientos en "Otoño" en los meses de septiembre, octubre, noviembre y "Primavera" en los meses de marzo, abril y mayo. Por lo tanto, también se buscará analizar si de los contenidos de EE. UU hay un aumento en esas épocas del año.

Debido al auge que están teniendo las plataformas de streaming en los últimos años analizaremos los contenidos de una de las plataformas más conocidas a nivel mundial como es Netflix que ofrece un amplio catálogo de contenidos audiovisuales. Según un informe de Netflix [3] del 2017 se plantea que el 50 % de su catálogo sea contenido original (procedente de EE. UU) y que al menos se produzcan 80 películas originales y 30 series animadas (género animación) para 2019 [2017 – 2019]. A continuación, comprobaremos si estas metas se cumplen con nuestros datos.

Por último, se pretende analizar cómo ha evolucionado en los últimos 20 años la producción de los contenidos que han ganado algún Oscar. Cabe destacar que existen diferentes premios Oscar como "Best Picture", "Best Director", etc. Los premios Oscar es un premio exclusivo de películas.

1. **Análisis de las producciones de cine nacional**

KPI 1.1: Comprobar si el crecimiento de películas españolas aumenta en un +30% del 2021 al 2024.

KPI 1.2: Comprobar si el crecimiento de series españolas aumenta en un +30% del 2021 al 2024.

2. **Análisis de las producciones más populares**

KPI 2.1: Analizar si las películas populares no han bajado su producción en un -15% en los últimos 20 años [2004 – 2024].

KPI 2.2: Analizar si pertenecen al género "Drama" al menos un 10% de las series populares en los últimos 20 años [2004 – 2024].

3. **Análisis de los lanzamientos en fechas clave de España y EE. UU**

KPI 3.1: Analizar que los contenidos de España lanzados en "Navidad" aumentan +15% en los últimos 20 años [2004 – 2024].

KPI 3.2: Analizar que los contenidos de España lanzados en "Verano" aumentan +20% en los últimos 20 años [2004 – 2024].

KPI 3.3: Analizar que los contenidos de EE. UU lanzados en "Otoño" aumentan +20% en los últimos 20 años [2004 – 2024].

KPI 3.4: Analizar que los contenidos de EE. UU lanzados en "Primavera" aumentan +20% en los últimos 20 años [2004 – 2024].

4. Análisis de los contenidos de Netflix

KPI 4.1: Analizar que el 50% de los contenidos de Netflix sea contenido procedente de EE. UU de 2017 a 2019.

KPI 4.2: Analizar que en total en Netflix se producen al menos 80 películas procedente de EE. UU de 2017 a 2019.

KPI 4.3: Analizar que en total en Netflix se producen al menos 30 series animadas de 2017 a 2019.

5. Análisis del reconocimiento en premios Oscar

KPI 5.1: Analizar si los contenidos premiados con Oscar aumentan en 5 unidades en los últimos 20 años. [2004 – 2024]

KPI 5.2: Analizar si en total hay al menos 100 contenidos que han ganado algún premio Oscar en los últimos 20 años [2004 – 2024].

Además, el sistema debe de ser capaz de responder a una serie de preguntas analíticas. Estas preguntas analíticas nos ayudaran a profundizar y detectar tendencias y patrones tanto históricos, como actuales y están relacionadas tanto con el ámbito nacional e internacional.
Algunas de estas preguntas son:

- ¿Qué géneros son los más dominantes?
- ¿Qué películas tienen una mayor recaudación?
- ¿En qué épocas se realiza más contenido?
- ¿De qué países son los actores más populares (estrellas consagradas)?
- ¿Qué idiomas son los más frecuentes en las películas?
- ¿Qué plataformas tienen contenidos mejores valorados?

- ¿A qué grupo de edad están los actores más populares en la actualidad?

Por último, de manera adicional el sistema debe ser escalable, para ello se utilizarán diferentes herramientas y tecnologías Big Data que se comentarán más adelante.
Además, el sistema debe estar preparado para poder incorporar nuevas herramientas que permitan integrar datos en tiempo real, asimismo como añadir más nodos de procesamiento para ser más tolerante a fallos.

4. Estado del arte

4.1 Evolución histórica y relevancia del análisis audiovisual

El análisis de contenidos audiovisuales es un tema que a día de hoy continúa siendo muy relevante. Desde finales del siglo XIX, el sector audiovisual ha sido fundamental para el entretenimiento y la cultura de la sociedad. Un momento muy importante en su historia fue la primera proyección pública de los hermanos Lumière en el año 1895, este hecho marcó el nacimiento de la industria cinematográfica. El cine ha evolucionado significativamente con el paso del tiempo en diferentes aspectos, como en el uso de nuevas tecnologías, narrativa y la forma que se analiza su impacto y relevancia.

El cine históricamente ha sido y es una de las industrias con mayor impacto y alcance a nivel mundial. Su importancia no solo se encuentra en el entretenimiento, sino que también está en la diversidad de géneros que ofrece, lo que permite atraer a un público con diferentes gustos y diferentes edades.

Por otro lado, las películas muestran la sociedad dando una visión de diferentes épocas, clases sociales y culturas. Las películas pueden ofrecer información importante como la vida de personajes históricos.

Cada película involucra a un grupo de personas, como actores, directores y guionistas, los cuales tienen un papel crucial para determinar si una película será exitosa o no.

Con el avance de las nuevas tecnologías, el análisis de películas ha progresado de manera notable. La capacidad de trabajar con grandes volúmenes de datos está permitiendo a las personas que trabajan en este sector, identificar patrones y tendencias que años atrás eran indetectables. Algunos de estos patrones son: preferencias de la audiencia, tendencias en la taquilla y el impacto de factores sociales y económicos en el éxito de una película.

El mercado audiovisual, es uno de los más importantes y que tiene un mayor crecimiento a nivel mundial. En el año 2020 la industria cinematográfica a nivel mundial generó unos ingresos superiores a 42.000 millones de dólares. Esto se muestra en el informe de la Motion Picture Association (MPA, 2021) [3]. Con esto se puede observar el gran impacto que tiene el mundo del cine en diferentes países, con mercados importantes como China, EE. UU e India que lideran la producción y distribución de películas [4].

El uso de tecnologías modernas como el Big Data a la hora de analizar películas ofrece diversas ventajas tanto para las diferentes compañías cinematográficas como para los espectadores. La principal ventaja es la posibilidad de comprender mejor las preferencias del público, esto permite a los trabajadores de la industria cinematográfica adaptar el contenido audiovisual a los intereses de la audiencia.

4.2 Herramientas y plataformas de análisis audiovisual

Es importante destacar que el papel de las plataformas de streaming (como Netflix, Amazon Prime Video, Disney+), las cuales han cambiado por completo la forma en la que se consume el contenido audiovisual. Muchas plataformas de streaming utilizan tecnologías para analizar datos de sus usuarios. Esta información les permite optimizar sus catálogos, recomendar contenido personalizado, y ayudara en la toma de decisiones sobre que contenido audiovisual adquirir y cual no. Con el análisis de datos que realiza las plataformas de streaming, pueden mejorar la experiencia del usuario ajustándose a sus preferencias.

Además, existen diferentes páginas webs que nos ofrecen diferentes datos de una película o serie, como las personas involucradas en ella, fecha de estreno, recaudación, etc. Algunas de estas páginas webs son IMDb, JustWatch o Filmaffinity.

No obstante, este tipo de páginas están diseñadas para que el usuario pueda realizar consultas individuales, es decir consultar la información de manera individual de un contenido o actor. Este tipo de páginas, no están diseñadas para ofrecer un servicio de análisis con una cantidad de datos masiva en donde se puede analizar diferentes aspectos de los contenidos audiovisuales, esto sí que lo ofrece este proyecto.

La principal limitación que tienen este tipo de webs es que no permiten realizar análisis personalizados ni hacer una relación entre datos como la puntuación y la recaudación. En este aspecto este proyecto se diferencia en que tiene un enfoque más analítico, que permite cruzar diferentes datos y responder a ciertas cuestiones analíticas que contestaremos más adelante. El proyecto se centra más en el tratamiento masivo de datos, mientras que las otras plataformas son para realizar consultas individuales

4.3 Introducción al Business Intelligence

El origen del Business Intelligencie (BI), comienza en una época en donde las empresas dependían del almacenamiento físico de documentos, almacenando la información en papel y también realizaban análisis manuales con la ayuda de hojas de cálculo. Gracias al avance de la tecnología, se comienza a usar bases de datos y sistemas para gestionar la información más avanzados. Durante las décadas de 1950 y 1960, grandes empresas como IBM, introdujeron sistemas de procesamiento de datos. No obstante, no fue hasta la década de 1980, que el concepto Business Intelligence se comenzó a usar. Este término fue inventado por Howard Dresner de Gartner [5].

4.3.1 Definición Business Intelligence

La Inteligencia de Negocio o Business Intelligence (BI) son todas las tecnologías con las cuales se puede recopilar y analizar datos, los cuales serán de gran utilidad a las empresas a la hora de tomar decisiones basándose en los datos recogidos. Dichos datos pueden ser usados en diferentes áreas y adaptarlos para analizarlos posteriormente, con esto se consigue mejorar la eficiencia de los procesos y de amoldarse a diversas situaciones en el mercado [6].

Para conseguir obtener información de relevancia en un sistema de BI, es fundamental recopilar datos de diferentes fuentes que pueden ser tanto fuentes externas como internas. Los datos se pueden clasificar en datos estructurados, que son aquellos que se organizan en bases de datos y no estructurados, como los emails, publicaciones en redes sociales o documentos. No obstante, uno de los mayores retos que surgen dentro del BI es la información que se ha obtenido, integrarla de forma eficiente, ya que se pueden encontrar diferentes estructuras y formatos.

En este aspecto, en los procesos ETLs (Extract, Transform and Load), es fundamental asegurarse que os datos sean consistentes y estén preparados para proceder con su análisis. De hecho, se estima que la mayor parte del esfuerzo en proyectos que usan BI, aproximadamente un 70% y 80% del esfuerzo en proyectos BI, es destinado a limpieza e integración de datos antes de ser usados por empresas a la hora de toma de decisiones. Las empresas afrontan problemas como la calidad de los datos, compatibilidad de sistemas, etc. En cierta medida, el éxito de un proyecto de BI depende que los datos se integren de forma que los mismos sean coherentes y estén actualizados [7].

4.3.2 Importancia y ventajas del BI

El Business Intelligence es una herramienta fundamental para las empresas porque permite convertir grandes volúmenes de datos en información útil a la hora de tomar decisiones. Su uso tiene numerosas ventajas [8]:

- **Ventaja competitiva**: Cuando se analiza el comportamiento del mercado, las empresas son capaces de detectar patrones de consumo lo cual les permite adelantarse a su competencia. Un ejemplo sería una plataforma de streaming como Netflix, la cual puede identificar que géneros gustan más a sus usuarios y priorizar la compra de contenido que pertenezcan a estos géneros. Lo que permitirá satisfacer la demanda de sus usuarios.
- **Decisiones más eficientes**: La posibilidad de acceder a los datos de manera precisa, permite evaluar el posible impacto de estrategias antes de su puesta en marcha. Por ejemplo, una productora puede analizar datos sobre el éxito de películas parecidas en diferentes mercados para deducir en que países invertir en distribución o que estrategias de marketing utilizar.
- **Eficiencia en producción**: La optimización de recursos es un aspecto fundamental en el sector audiovisual. Analizando el coste de rodaje, tiempo de la producción y recaudación de la producción audiovisual en la taquilla o plataformas de streaming, las compañías productoras son capaces de detectar áreas donde reducir gastos sin poner en riesgo la calidad del contenido audiovisual. Un ejemplo es ajustar los lugares del rodaje o ajustar los días, para evitar gastos innecesarios.
- **Cumplimiento de normativas:** La industria cinematográfica es un área en donde los derechos de autor es un aspecto clave para tener en cuenta. Con el uso de herramientas BI, se puede realizar supervisión de licencias y derechos de distribución. Por ejemplo, si una plataforma de streaming planea extenderse a un nuevo país en el cual no está disponible, deberá analizar datos para comprobar que su contenido cumpa con las normativas y licencias presentes en ese país.

4.4 Arquitectura BI

La arquitectura BI está compuesta por varios procesos. Entre los cuales se incluyen las fuentes de datos, los procesos ETL, el almacén de datos, el servidor OLAP y el dashboard Los cuales serán explicados más adelante.

Figura 6. Arquitectura BI - wordpress.com

4.4.1 Fuentes de datos

En primer lugar, en una arquitectura BI encontramos las fuentes de datos, de las cuales se extraen la información. Las fuentes de datos deben ser fiables para asegurar que los análisis realizados estén lo más actualizado posible y se realicen de forma eficiente. Estas incluyen bases de datos, archivos de texto, hojas de cálculo, APIs y servicios web, etc.

Como se dijo anteriormente, los datos pueden ser obtenidos de diferentes fuentes. Algunas de estas son:

- **BB.DD transaccionales:** En estas bases de datos se almacena información fundamental de la empresa, como ventas, stock, etc.
- **APIs y servicios web:** Permiten recoger información de fuentes externas en tiempo real, lo que asegura que la información recogida esté actualizada.
- **Ficheros transaccionales:** Algunas empresas hoy en día dependen de ficheros transaccionales los cuales pueden ser de aplicaciones tanto internas como externas. Estos ficheros contienen información generada operaciones diarias como ventas, registro de usuarios y se puede exportar a diferentes formatos. Algunos de los formatos de datos estructurados son:
    - **CSV:** Destaca por su compatibilidad y fácil integración a la hora de trabajar con este tipo de ficheros, aunque no soporta tipos de datos complejos.

- **Excel:** A día de hoy se sigue utilizando en el entorno empresarial para visualizar y manipular datos
- **JSON:** Utilizado en aplicaciones webs y APIs, nos permite organizar la información de forma jerárquica.

También cabe destacar el sistema de las 5 estrellas de calidad, en el cual se propone cinco niveles en el que nos permite evaluar la calidad, la accesibilidad y reutilización de datos. Cada nivel define un escalón de madurez a la hora que se publican los datos.

- ★ Los datos están publicados, con cualquier licencia en cualquier formato como JPG, PDF, etc.
- ★★ Los datos publicados, usan un formato estructurado, es propietario, lo que quiere decir que lo controla la empresa como las hojas de cálculo de Excel.
- ★★★ Los datos están en formato abierto, su especificación es pública y cualquier usuario puede implementarla, como CSV, JSON, etc.
- ★★★★ Los datos son asignados a URIs, estos son identificadores únicos, cada elemento puede referenciarse y lo podemos recuperar de forma individual mediante este identificador.
- ★★★★★ Los datos se enlazan entre sí con otros datasets como Linked Data, lo que nos facilita hacer consultas complejas.

Este se creó para ser usados con datos abiertos, pero también lo usan las empresas para valorar el nivel de reutilización de datos.

- **Aplicaciones internas:** Existen empresas que desarrollan sus propias aplicaciones para gestionar los datos ajustado a sus intereses. Por ejemplo, una productora cinematográfica implementa un sistema para registrar el desempeño de sus películas en distintas plataformas de streaming.

Un aspecto a tener en cuenta sobre los datos es la topología de los mismos, como se conectan y circulan dentro de la arquitectura BI. Un buen diseño de datos evitaría redundancias y aseguraría una correcta transformación de los datos, de manera que estos se almacenen de forma limpia y estructurada.

Cabe destacar, que no solo es suficiente con la recopilación de datos, también hay que tener en cuenta la calidad de los mismos. Ya que, si estos son erróneos, los análisis que se realicen

no serán correctos. Para medir la calidad de datos, se utilizan modelos como el de "Estrellas de Calidad", que mide parámetros como la consistencia y la precisión entre otros. Un ejemplo sería un cine que detecta duplicados en registros de películas en su base de datos, el análisis que harán para evaluar las películas será incorrecto. [9]

4.4.2 Procesos ETL

Una vez que se han identificado y analizado las diferentes fuentes de datos, el siguiente paso es definir los ETL los cuales son una estrategia de integración de la información, la cual puede provenir de diferentes fuentes. En la imagen anterior, aunque se muestre con unos simples engranajes es una de las partes o la parte más compleja de un sistema BI. Como dice su propio nombre, los ETL permiten extraer, transformar y cargar los datos a un destino final, como una base de datos. A pesar de que tradicionalmente, estos procesos se han gestionado con los ETL, existen otras alternativas como ELT.

En los procesos ETL, la primera fase como se dijo anteriormente es la de extracción, en la cual analizamos las fuentes de datos que ya se han identificado anteriormente, solventando posibles problemas como incompatibilidad de datos, descartando columnas no útiles y quedándonos con la información necesaria.

Lo siguiente es la transformación de los datos, o dicho de otra manera limpiar los datos, corregir datos incompletos, eliminar campos duplicados, etc. Esta fase es vital para evitar futuros problemas como puede ser la creación del dashboard erróneo por contener valores erróneos por los valores duplicados o por algunas medidas que no sean correctas.

La última fase es la de carga, en la cual decidiremos en que estructura se almacenarán los datos, que puede ser un almacén de datos. En esta fase, además de añadir los datos, se actualizarán los ya existentes.

Los ETLs son de utilidad cuando es necesario un control riguroso en la transformación de datos, anteriormente de que estos sean almacenados, además el proceso de limpieza se realiza de manera controlada.

Anteriormente se mencionó que también existen los ELTs, una alternativa a los ETLs. La diferencia principal es el orden que se realizan las operaciones. En conclusión, los ETLs son más útiles cuando se quiere con trabajar de datos de forma masiva, antes que estos sean almacenados, mientras que los ELTs destacan cuando disponemos de sistemas con gran

capacidad de procesar datos, permitiendo que estos primero se extraigan, se carguen y luego se transformen. Esto es ideal en proyectos de Big Data que trabajen con grandes volúmenes de datos que necesitan ser analizados en tiempo real [10].

4.4.3 Almacén de datos

El almacén de datos o Data Warehouse, es un elemento fundamental en las arquitecturas BI. Este consiste en un repositorio centralizado, el cual acumula información y permite a los usuarios un acceso fácil a los datos. La información acumulada proviene de diferentes fuentes tras ser extraídas por los ETL, como se explicó en el apartado anterior.

El concepto de almacén de datos surgió a finales de los años 80 y ha ido evolucionando constantemente. Cabe destacar dos nombres aquí: Bill Inmon y Ralph Kimball, quienes formularon diferentes enfoques para la creación del Data Warehouse.
En primer lugar, Inmon afirma que el Data Warehouse debe construirse a partir de datos transformados previamente. Inmon dice que la base de datos resultante debe cumplir con unos requisitos. Primero que la información se integre de diferentes fuentes, lo siguiente es que el almacén de datos sea no volátil, es decir que los datos una vez son almacenados estos no cambiaran. Además, debe guardar información a lo largo del tiempo para poder realizar comparaciones [11].

Por otro lado, Kimball defiende por iniciar el proceso mediante la creación de Data Marts, los cuales son subconjuntos que se especializan en un área específica, los cuales contienen grandes volúmenes de datos. Los Data Marts, están diseñados para responder a necesidades puntuales, y se integran posteriormente en el almacén de datos [12].

Kimball creó el **modelo multidimensional,** en el cual se organiza la información de manera que los datos se puede analizar con diferentes técnicas de diseño como el estrella o copo de nieve, para que sea más sencillo el análisis y la consulta de datos. El modelo multidimensional es fundamental en los Data Warehouses porque organiza y estructura los datos en dimensiones y hechos. Gracias a este modelo, los usuarios pueden analizar y consultar los datos de manera más sencilla, permitiendo visualizar la información desde varias perspectivas.

Este modelo está formado por tres partes:

- **Dimensiones**: Son los puntos de vista desde los cuales se pueden analizar los datos, algunos ejemplos serían el tiempo, la ubicación, el cliente o las ventas. Las dimensiones estarán compuestas por miembros, por ejemplo, zona y área son miembros de la dimensión de ventas. Además, los miembros normalmente se organizan en jerarquías.
- **Medidas**: Se definen como datos cuantitativos que analizarán, como ventas, ingresos, costos. Las medidas son datos numéricos que permiten la evaluación del rendimiento.
- **Hechos:** Los hechos constituyen el centro en el análisis porqué todas las dimensiones y jerarquías derivan de los hechos que forman el esquema estrella. Por ejemplo, si estamos analizando ventas en una tienda, el hecho podría ser la transacción.

A continuación, se muestra una imagen con un esquema estrella en donde la tabla de hechos es "Ventas" y las dimensiones son "Clientes", "Productos" y "Fechas". Esta estructura facilita el análisis de las ventas desde diferentes perspectivas, como quién compró, qué se compró y cuándo se compró.

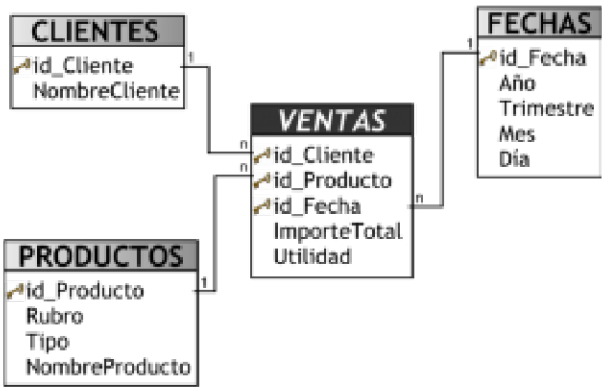


Figura 7. Esquema estrella – dataprix.com

Además del Data Warehouse, también cabe destacar que con el crecimiento de forma exponencial de datos y la gran variedad de fuentes existentes, han ido surgiendo nuevas estrategias para gestionar la información.

En primer lugar, encontramos los Data Lakes, estos son repositorios que permiten almacenar grandes volúmenes de datos en su formato inicial, estos fatos pueden ser tanto estructurados como no estructurados. La gran diferencia con el Data Warehouse es que no requiere la transformación de datos, lo que facilita que estos datos se añadan de manera rápida [13].

En el caso de que un Data Lake no este catalogado y no tenga un buen control de calidad puede dar lugar a lo que se conoce como Data Swamp. En este caso los datos serán inútiles porque dificulta el aprovechamiento de los mismos, generando problemas en la consistencia y en la seguridad de los datos.

También encontramos Data Mesh, en este caso se plantea una reorganización en la manera que se gestionan los datos. En vez de concentrar todo en un único sistema, se reparten los datos en diferentes equipos. De esta manera se fomenta la colaboración interna y posibilita que el acceso a la información sea más rápido. El Data Mesh es útil si se trabaja con grandes volúmenes de datos y fuentes variadas. [14]

4.4.4 Servidor de consultas (OLAP)

Los servidores OLAP (Online Analytical Processing), fueron diseñados para realizar consultas complejas y agilizar el análisis de grandes volúmenes de datos. A diferencia de las bases de datos transaccionales (OLTP), las cuales surgieron para introducir, actualizar o borrar datos, OLAP realiza consultas avanzadas y proporciona respuestas rápidas a preguntas analíticas [15].

En este caso se aprovecha la organización multidimensional de la información, lo cual permite a los usuarios examinarla desde diferentes ángulos. Por ejemplo, se pueden combinar variables como periodos de tiempo, ubicaciones, etc. También cabe añadir que las estructuras OLAP optimizan la ejecución de queries, ya que calculan los datos y posteriormente los almacenan, esto hace más pequeño el tiempo de respuesta y ayuda en la toma de decisiones.

En la siguiente figura, podemos observar un ejemplo de cubo OLAP, el cual muestra la forma en la que se distribuyen los datos en diferentes dimensiones. En primer lugar, encontramos la línea de productos, posteriormente las ventas y por último los trimestres o años. Cada celda del cubo representa un valor agregado, como por ejemplo el total en ventas o el promedio en costos, el cual es el resultado de la intersección de las dimensiones nombradas. La representación gráfica permite dividir el cubo para enfocarse en un área específica, lo que optimiza las consultas, permitiendo análisis detallados.

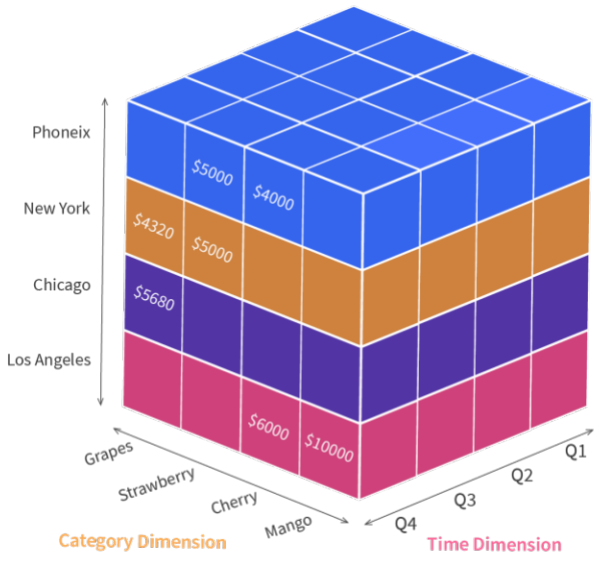


Figura 8. Ejemplo cubo OLAP - dba.stackexchange.com

4.4.5 Visualización o Dashboard

Por último, está la parte de visualización de datos, la cual permite visualizar los datos y la posibilidad de realizar consultas. Mediante cuadros de mandos o Dashboards, es posible representar datos complejos de manera clara y atractiva visualmente, lo que facilita identificar patrones y desviaciones en los indicadores clave de rendimiento conocidos como KPIs.

Los Dashboards son importantes ya que ayudan a la toma de decisiones en una empresa. Un ejemplo sería un Dashboard que muestra ventas mensuales realizadas y las compara con las ventas esperadas.

En la siguiente figura observamos un ejemplo de Dashboard que muestra diferentes gráficos para ofrecer una visión más detallada del desempeño de un negocio.

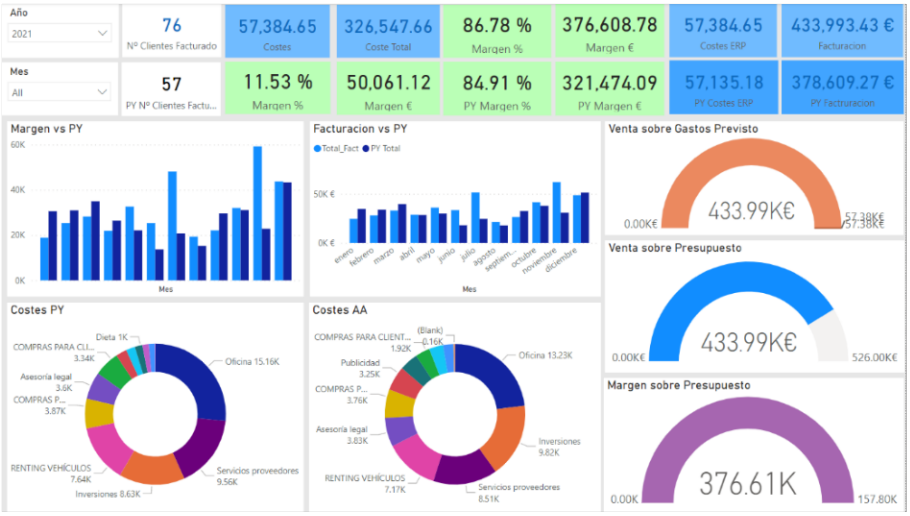


Figura 9. Ejemplo Dashboard – evotic.es

En resumen, la visualización de datos nos ayuda a identificar patrones y también para la toma de decisiones. Existen diversas herramientas para la visualización de datos, algunas de ellas son: Power BI, Tableau y Qlik Sense.

4.5 Problemas de la arquitectura BI

Aunque la arquitectura del BI es bastante robusta, tiene algunos inconvenientes. Uno de los más destacados es la dificultad a la hora de integrar datos. Las empresas suelen manejar información que se encuentra en diferentes ubicaciones y formatos, esto complica mucho trasladar los datos al Data Warehouse y como consecuencia puede generar datos inconsistentes y afectar negativamente los análisis.

Otro problema importante es el tiempo y el costo que implica implementar este tipo de arquitectura. Además de requerir bastante inversión en software, hardware y personal, también es necesario adaptar la arquitectura a las necesidades específicas de la empresa y mantenerla operativa.

Cabe destacar que la escalabilidad también es un problema, porque a medida que el volumen de datos crece, los tiempos de respuesta pueden volverse más lentos y esto provoca una bajada en el rendimiento general.

Además, la falta de flexibilidad es otra desventaja. Las soluciones tradicionales de BI no se ajustan fácilmente a los cambios rápidos en los requisitos del negocio o a la incorporación

de nuevas fuentes de datos, lo que limita la capacidad de una empresa para reaccionar rápido ante nuevas oportunidades.

Por último, la privacidad y la seguridad de los datos son otros problemas importantes. Las arquitecturas BI suelen manejar grandes volúmenes de información sensible, lo que aumenta los riesgos de seguridad. Por eso, es esencial que las empresas tomen medidas para proteger esa información y evitar accesos no autorizados.

4.6 Historia del Big Data

Big Data como se dijo anteriormente, se puede definir como conjuntos de datos de gran tamaño y complejos los cuales no pueden ser trabajados con herramientas tradicionales de procesamiento de datos. Esto conlleva al uso de nuevas tecnologías para trabajar con grandes volúmenes de datos. Existen tres V que caracterizan al Big Data:

- **Volumen:** Cantidad de datos que se han almacenado
- **Variedad:** Diversidad en los tipos de datos ya que pueden ser en forma de imagen, texto, video, etc.
- **Velocidad:** Rapidez con la que se crean y procesan los datos.

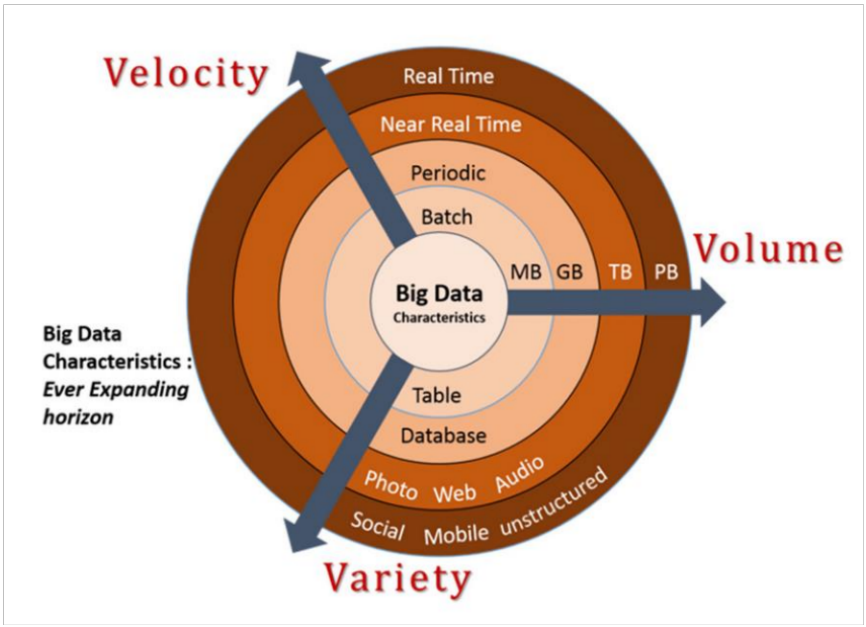


Figura 10. 3V Big Data - researchgate.net

El concepto Big Data ha evolucionado con el paso de los años.

Durante los años 90, con el crecimiento de Internet y la adopción de computadoras personales aumentaron considerablemente la cantidad de datos generados. Por lo tanto, surgieron nuevas tecnologías como los sistemas de procesamiento analítico en línea (OLAP) y los almacenes de datos, ya explicados anteriormente, para manejar grandes volúmenes de datos.

Durante los 2000, el concepto Big Data comenzó a hacerse popular. Surgieron nuevas herramientas para procesar datos como Hadoop, las cuales hicieron más fácil en manejo de grandes volúmenes de datos.

A partir de la década de 2010, el uso de tecnologías Big Data se trasladó a diferentes industrias. Surgieron nuevas herramientas como Apache Spark y Apache Flink, las cuales mejoraron la eficiencia a la hora de procesar datos. Además, otras áreas como la Inteligencia Artificial (IA) y el Machine Learnig (ML) complementaron al Big Data para extraer datos de manera más precisa.

4.7 Tecnologías Big Data

Como se ha mencionado anteriormente la variedad, el volumen y la velocidad de los datos (3V) creados han aumentado de manera exponencial. Esto ha provocado el nacimiento de nuevas tecnologías para analizar y gestionar Big Data. Estas tecnologías abarcan un gran número de herramientas para procesar datos de manera eficiente, almacenarlos y visualizarlos. A continuación, se explicarán algunas de esas herramientas.

Además, cabe destacar que, a lo largo del tiempo, se propusieron expandir las 3V originales de Big Data para añadir nuevas características como la veracidad que hace referencia a la calidad y fiabilidad de la información y la variabilidad. En este caso estaríamos hablando de 5V. Estas nuevas dimensiones, permiten entender mejor las oportunidades relacionadas al manejo de grandes volúmenes de datos [16].

4.7.1 Landscape Big Data

El Landscape se puede definir como el conjunto de tecnologías y herramientas que se han desarrollado para trabajar con grandes volúmenes de datos, tratando temas como procesamiento, almacenamiento y análisis de datos.

Con el paso de los años, el ecosistema va creciendo de manera exponencial para abarcar los temas mencionados anteriormente. Como se explicó en puntos anteriores, en el landscape actual encontramos tecnologías modernas como Data Lakes, IA, Machine Learning procesamiento en tiempo real, etc.

En la siguiente figura, podemos observar un esquema actualizado del Landscape de Big Data, en el cual se muestra de manera gráfica las herramientas que componen este ecosistema [17].

En puntos posteriores hablaremos sobre algunas de las tecnologías.

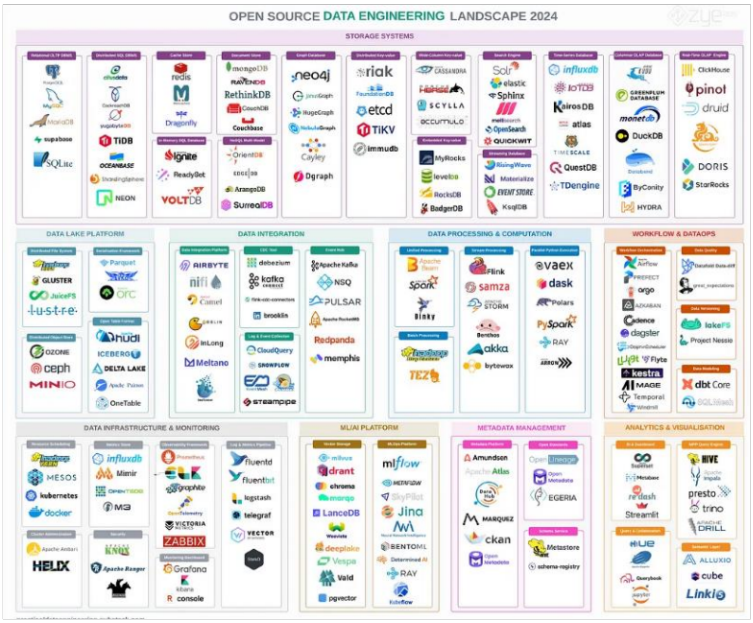


Figura 11. Big Data Landscape 2024 - pracdata.io

4.7.2 Herramientas de procesamiento

4.7.2.2 Spark

Apache Spark es una herramienta de código abierto diseñado para procesar datos de manera rápida y fácil. Se conoce por su capacidad para realizar análisis complejos y gestionar

proyectos de análisis en los que el tiempo es crítico. Además, nos proporciona una interfaz unificada que permite organizar clústeres y también es tolerante a fallos.

Spark destaca por su capacidad de procesamiento en memoria. También admite una variedad de lenguajes de programación como Java, Scala, Python y R, etc. Por otra parte, su API está diseñada para ser fácil de usar, lo que nos permite escribir programas en menos líneas de código. Además, Spark incluye diferentes librerías. A continuación, podemos ver una imagen en las que se muestran dichas librerías y posteriormente las explicaremos.

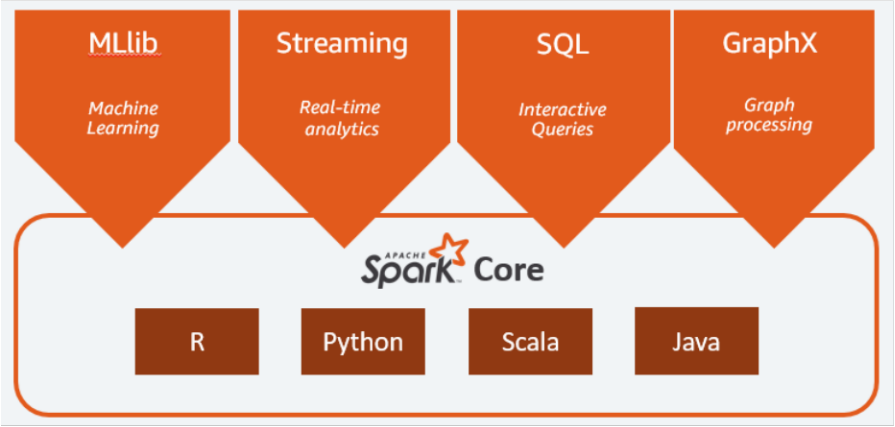


Figura 12. Librerías Spark - medium.com

Además, cabe destacar que Spark es una herramienta muy rápida, esto se debe a su capacidad para realizar operaciones en memoria, lo nos permite reducir tiempos de lectura y escritura en disco. Esto es útil para aplicaciones dinámicas, como el aprendizaje automático, que requieren múltiples pases de los mismos datos. Como ya dijimos, Spark también es escalable y puede usarse en entornos desde pequeños clústeres hasta grandes centros de datos. En resumen, Spark nos permite realizar tareas analíticas, como consultas SQL y análisis de gráficos, todo en una única plataforma [18].

4.7.2.2.1 Spark MLlib

Como vimos en la imagen anterior Apache Spark cuenta con diferentes librerías, una de ellas es MLlib, la cual es una librería de Machine Learning. Mlib fue diseñada para poder integrar el Machine Learning y de garantizar que sea usado de manera sencilla. Para ello MLIb proporciona diferentes herramientas para realizar algunas tareas de Machine Learning.

Cabe destacar que MLlib nos proporciona un conjunto de algoritmos de Machine Learning para utilizar el procesador distribuido de Spark. Estos algoritmos encontramos la regresión logística, métodos de regresión como por ejemplo la regresión lineal y algoritmos de

clustering como K-means. Además, MLlib admite la reducción de dimensionalidad mediante algoritmos como PCA y SVD, así como la extracción de características y la transformación de datos.

También podemos crear pipelines de Machine Learning, lo que ayuda a la creación y el mantenimiento de flujos de trabajo de aprendizaje automático. Los pipelines combinan múltiples procesos de transformación de datos y aprendizaje automático en una serie de pasos secuenciales que permiten un procesamiento y una configuración de parámetros más eficientes [19].

4.7.2.2.2 Spark Streaming

Spark Streaming es otra librería de Apache Spark que nos permite procesar datos en tiempo real. Nos proporciona una API de alto nivel la cual facilita trabajar de forma continua. Además, los datos provienen de diferentes fuentes como Kafka y sockets TCP, los cuales nos permite realizar diferentes operaciones en tiempo real.

En vez de procesar cada dato a medida que llega, Spark Streaming divide el flujo en pequeñas partes que se procesan con el motor distribuido de Spark, simplificando considerablemente la gestión de grandes volúmenes de datos.

Además, Spark Streaming nos proporciona una API unificada la cual soporta tanto el procesamiento tanto en tiempo real como en pequeñas partes, esto nos facilita el desarrollo de aplicaciones híbridas que requieren ambos tipos de procesamiento [20].

4.7.2.2.3 Spark SQL

Otra librería de Spark es Spark SQL, diseñada para trabajar con datos estructurados. Nos ofrece una interfaz que permite realizar consultas SQL incluyendo las funciones de procesamiento de datos de Spark. Esto nos ayuda a integrar SQL en el análisis de Big Data aprovechando las ventajas de Spark mediante lenguaje SQL [21].

Algunas de las características principales de Spark SQL son:

- **Consultas SQL y API de Dataset**: Nos permite ejecutar consultas SQL y utilizar la API de Dataset, que combina las ventajas de la programación orientada a objetos para poder realizar consultas SQL.

- **Catalyst Optimizer**: Incorpora un optimizador de consultas el cual analiza y nos permite mejorar las operaciones SQL y datos para lograr un rendimiento más eficiente.

- **Diferentes fuentes de datos**: Podemos leer la información desde diferentes fuentes y diferentes archivos como por ejemplo JSON, ORC, HDFS, etc. Además, podemos obtenerlos de diferentes sistemas en la nueve y bases de datos como MySQL.

- **Integración con Hive**: Es compatible con los metadatos de Hive, lo que nos permite realizar consultas en Hive sin necesidad de tener que modificar el esquema de datos ya existente.

4.7.2.2.4 Spark GraphX

Por último, encontramos la librería Spark GraphX, la cual se utiliza para el procesamiento y análisis de grafos a gran escala. Esta, combina las capacidades de procesamiento de datos de Spark con su habilidad para manejar grafos, ofreciéndonos una API unificada la cual nos permite trabajar tanto con datos como con grafos [22].

Algunas de sus principales características que encontramos son:

- **API Unificada**: GraphX como dijimos, nos ofrece una interfaz única para manipular grafos, esto nos permite ejecutar algoritmos de grafos en la misma plataforma que realizamos las operaciones, como consultas SQL.

- **Modelos de Grafo Flexibles**: GraphX utiliza un modelo basado en propiedades, esto quiere decir que cada vértice y artista pueden tener atributos asociados. Lo que nos proporciona una mayor flexibilidad para representar y manipular datos.

- **Algoritmos de Grafos**: No es necesario implementemos los algoritmos desde cero, porque GraphX incluye una biblioteca con algoritmos ya definidos como PageRank, Triangle Count, entre otros, esto nos facilita el análisis de grafos complejos.

4.7.2.3 Databricks

Databricks es una plataforma de procesamiento de datos en la nube basada en Apache Spark, la cual acelera el desarrollo de soluciones de datos y Machine Learning en la nube. Los databricks, proporcionan un espacio de trabajo colaborativo, en donde varios usuarios pueden trabajar en un único entorno compartido simultáneamente, pueden editar notebooks a la vez, realizar un control de versiones mediante Git y programar su ejecución como Jobs periódicos.

Una de sus principales ventajas es la gestión automática de los clústeres, ya que de manera sencilla se puede crear un clúster escalable el cual se puede ajustar a la carga de trabajo, puede detener o iniciar nodos, lo que ayuda a la optimización de recursos y costes [23].

Para el almacenamiento y la seguridad de datos, Databricks se basa en Delta Lake, un formato transaccional que añade operaciones ACID sobre el Data Lake y permitirá consultar versiones anteriores de los datos, lo cual ayuda mucho en la tarea de recuperación ante errores.

Los Databricks incorporan Photon, el cual es su propio motor de ejecución optimizado en C++ que acelera consultas SQL y transformaciones de Data Frames, lo que reduce considerablemente los tiempos de consulta en comparación con Spark.

4.7.3 Herramientas de streaming y mensajería

4.7.3.1 Apache Flink

Apache Flink es un software que podemos usar para procesar datos, está diseñado para análisis en tiempo real (streaming). Puede manejar grandes volúmenes de datos, lo que lo convierte en una alternativa a Spark, con ventajas como la baja latencia y un rendimiento continuo.

Cabe destacar que una de las grandes ventajas de Flink es como dijimos el procesamiento en tiempo real con muy poca latencia, esto es muy útil para aplicaciones las cuales requieren una respuesta inmediata. Además, su arquitectura permite el procesamiento de datos en lotes y también en streaming, lo que nos facilita gestionar diferentes tareas. También es tolerante a fallos, lo que nos garantiza una alta disponibilidad del sistema [24].

Centrándonos más en la arquitectura, Flink está formado por los siguientes componentes:

- **Job Manager:** Se encarga de organizar la ejecución de tareas, además de asignar trabajos a los Task Managers, también monitorea operaciones y se encarga de la recuperación ante posibles fallos.
- **Task Managers:** Se encarga de ejecutar una parte del trabajo, lo que quiere decir que gestiona una fracción de datos.
- **Client:** Es la interfaz por la cual enviamos trabajos a Job Manager ya sea mediante comandos o una API.
- **Flink Runtime:** Es el núcleo del sistema, ejecuta tareas de procesamiento en tiempo real.
- **State Backends:** Se encargan de controlar el estado de las aplicaciones, esto es muy importante para tener la posibilidad de recuperación ante errores a la hora de procesar datos.

A continuación, en la siguiente figura podemos ver la arquitectura de Apache Flink, en la cual se muestran los componentes de los cuales se han mencionado anteriormente.

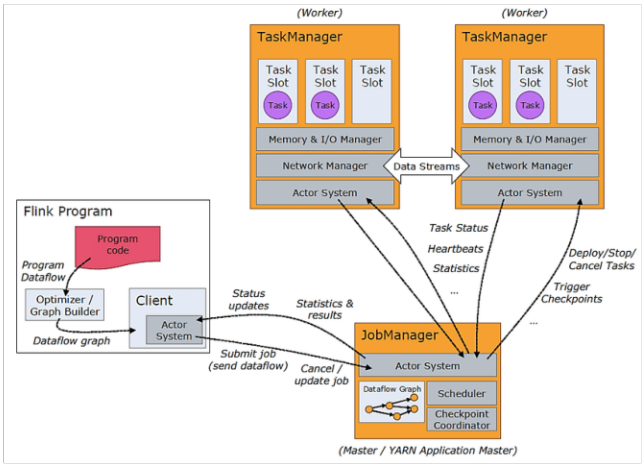


Figura 13. Arquitectura Apache Flink - apache.org

Hoy en día, además de Flink y Spark existen otras alternativas interesantes en el área del Big Data. Un ejemplo sería Apache Beam la cual permite definir pipelines de procesamiento tanto en lotes como en streaming. Lo que marca la diferencia de Beam es que se puede ejecutar sobre diferentes motores como Flink o Spark, esto nos proporciona una mayor flexibilidad [25].

Cabe destacar, que otra herramienta es Apache Iceberg, esta tiene un formato de tabla de código abierto diseñado para gestionar datos en Data Lakes. Iceberg nos permite trabajar con

datos de manera eficiente, ya que destaca por tener algunas características como transacciones ACID, manejo de versiones, integración de datos de manera optimizada acoplada a Spark o Flink. Esto nos ayudaría a almacenar los datos de manera consistente [26].

4.7.3.2 Apache Kafka

Apache Kafka es otra herramienta la cual podemos usar para transmisión de eventos, esta fue diseñada para crear aplicaciones de datos en tiempo real. Fue creado por LinkedIn en sus inicios y posteriormente fue donado a Apache Software Foundation. Con el paso del tiempo se ha ido convirtiendo en una herramienta importante para gestionar datos en tiempo real [27].

Entre las características de Kafka, podemos destacar:

- **Alto rendimiento:** Kafka es capaz de manejar un número muy alto de eventos por segundo, lo que lo hace ideal para aplicaciones que requieren un alto rendimiento de datos.
- **Escalabilidad**: Kafka se pude expandir horizontalmente, permite añadir nodos al cluster y esto nos permite gestionar mayores volúmenes de datos.
- **Durabilidad y persistencia:** Los datos se escriben en disco y se replican en varios nodos, esto nos garantiza que la información esté disponible incluso si hay fallos.
- **Flexibilidad en el consumo de datos:** Varios consumidores pueden leer la información a su ritmo, lo que facilita el desarrollo de aplicaciones complejas las cuales operan en tiempo real.
- **Integración:** Kafka se integra fácilmente con otras tecnologías de Big Data como Apache Hadoop, Apache Flink y Apache Spark.

En cuanto a la arquitectura, Apache Kafka consta de los siguientes componentes:

- **Producers:** Son las aplicaciones que publican mensajes en uno o más topics dentro del sistema.
- **Consumers:** Son programas que leen y procesan mensajes de programas Kafka.
- **Brokers:** Son los servidores que forman el clúster de Kafka. El broker recibe mensajes de los producers, los almacena en disco y los envía a los consumidores.

- **Topics:** Estas son las secciones a partir de las cuales se generan los mensajes. Cada topic se divide en secciones que permiten el procesamiento paralelo.
- **Zookeeper:** Es un sistema de coordinación distribuido que Kafka utiliza para gestionar la configuración y coordinar las operaciones del clúster.

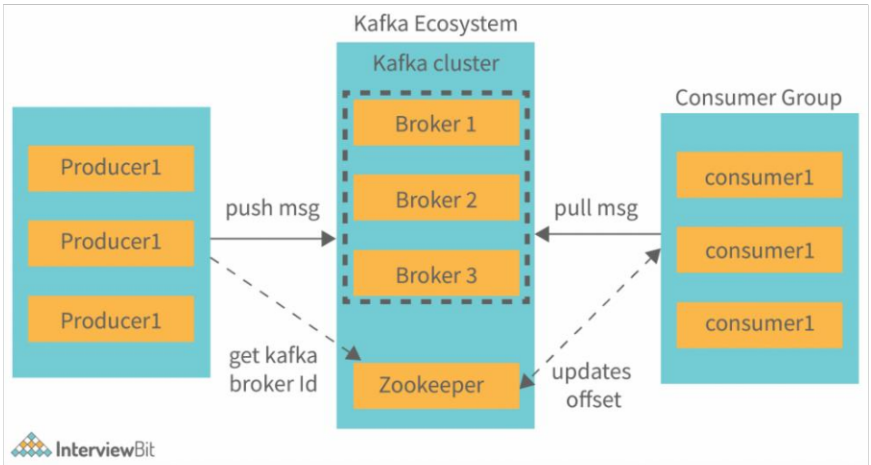


Figura 14. Arquitectura Apache Kafka - interviewbit.com

En resumen, Kafka es una buena opción si queremos transmitir eventos en tiempo real, además de que puede trabajar con grandes volúmenes de datos y se integra de manera sencilla con otras tecnologías.

4.7.4 Herramientas de almacenamiento y BD

4.7.4.1 Hive

Apache Hive lo podemos definir como una herramienta de Data Warehousing y análisis de datos la cual se ejecuta sobre Hadoop. Con Hive los usuarios harán consultas y pueden analizar grandes conjuntos de información almacenados en HDFS utilizando un lenguaje similar a SQL conocido como HiveQL.

En sus inicios, Hive fue desarrollado por Facebook para ayudar a los ingenieros a gestionar y consultar datos útiles para sus usuarios. Antes de que existiera Hive, el procesamiento de datos en Hadoop requería escribir complejos programas en MapReduce, lo cual era muy lento. Hive fue creado para proporcionar una interfaz más fácil y amigable para sus usuarios.

Si nos referimos a la arquitectura, Hive se basa en el cliente-servidor, donde el cliente envía consultas al servidor Hive, el cual luego las convierte en tareas de MapReduce que se ejecutan en Hadoop. La arquitectura de Hive se compone de varios componentes claves:

- **Metastore**: Se encarga de almacenar ña definición de tablas y esquemas de datos, es fundamental para la organización.
- **Driver**: Se encarga de revisar el tiempo de una consulta.
- **Compiler**: Transforma las consultas HiveQL en planes de ejecución de MapReduce.
- **Execution Engine**: Ejecuta los planes de ejecución de MapReduce generados por el Compiler.
- **HiveServer**: Nos ofrece una interfaz de usuario para ejecutar consultas y obtener resultados.

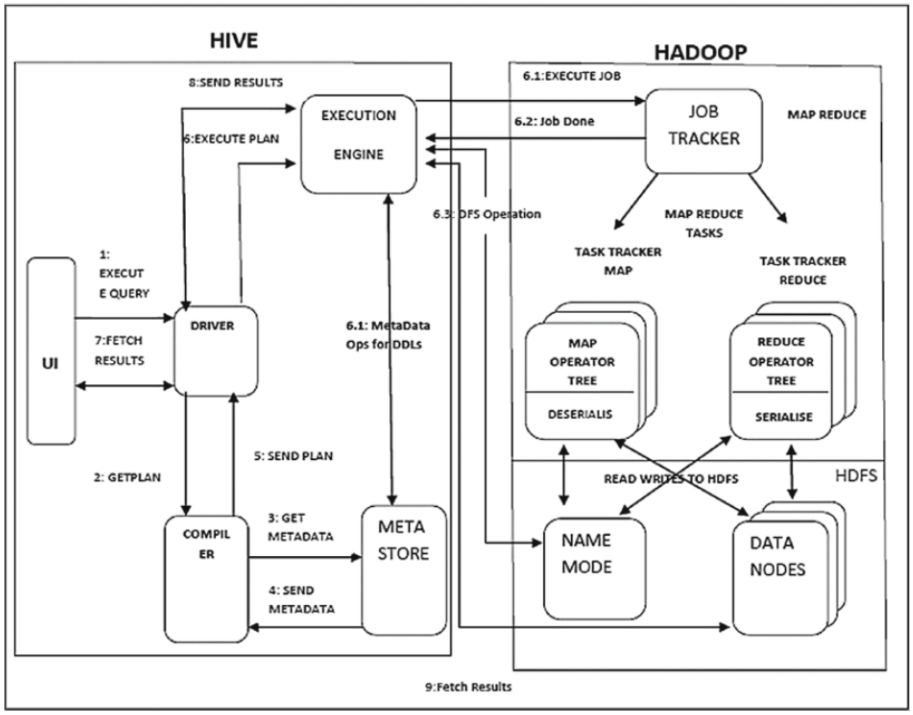


Figura 15. Arquitectura Apache Hive. forum.huawei.com

Por otro lado, cuando hablamos de la estructura de Hive encontramos las siguientes partes:

- **Tablas**: Forman la estructura básica y las podemos encontrar en diferentes formatos como texto, OCR, etc.
- **Bases de datos**: Nos permite organizar las tablas en grupos lógicos, facilitando administrarlas si encontramos un gran número de tablas.

- **Particiones**: Nos ayuda a dividir tablas de gran tamaño en grupos más pequeños, esto nos ayuda a mejorar la eficiencia a la hora de realizar consultas. Las particiones se basan en una o más columnas de la tabla.
- **Buckets**: Por último, las tablas las podemos dividir además en buckets, que los podemos definir como una especie de fragmentos de datos, para mejorar el rendimiento de las consultas JOIN.

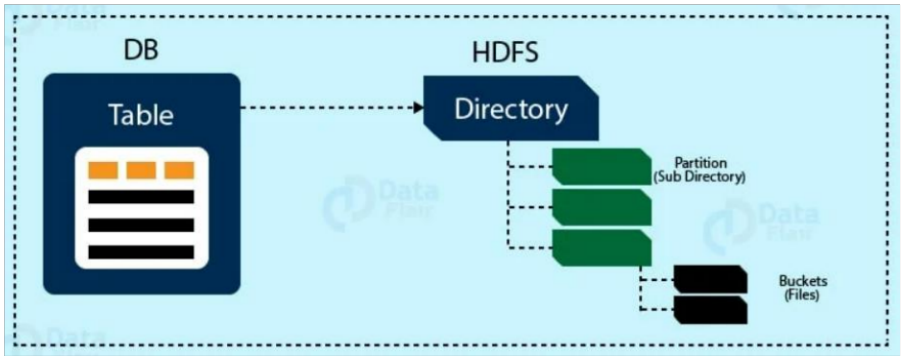


Figura 16. Estructura Apache Hive - developer.aliyun.com

En resumen, Hive nos puede ser útil para un Data Warehouse tradicional ya que ayuda a la gestión y consultas de datos los cuales están almacenados en HDFS. Además de que dispone de interfaz gráfica visualmente para que los usuarios puedan ver los datos de manera clara. También las consultas que se realizan son eficientes y efectivas como ya explicamos anteriormente [28].

Por último, cabe destacar que, a diferencia de otros sistemas para gestionar bases de datos, Apache Hive tiene una particularidad y es que no permite definir claves ajenas. Lo cual nos indica que la integridad referencial no la podemos gestionar de manera automática, sino que debemos controlarla mediante la lógica de la aplicación.

4.7.4.2 Cassandra

Apache Cassandra nació en Facebook. Su función fue mejorar el rendimiento del motor de búsqueda en la bandeja de entrada. Posteriormente, en 2008 se convirtió en un proyecto de código abierto y en 2010 fue adoptado como proyecto de alto nivel por la Apache Software Foundation en 2010.

Cassandra es una base de datos NOSQL, que está diseñada para trabajar con grandes volúmenes de datos.

Entre sus características encontramos:

- **Alta disponibilidad y tolerancia a fallos:** Su arquitectura nos asegura que la información continue accesible incluso si se producen errores, gracias a la replicación de datos.
- **Escalabilidad lineal:** Se puede agregar nuevos nodos al clúster sin interrumpor el servicio, esto nos permite gestionar de manera eficiente los datos.
- **Modelo de consistencia flexible:** Los usuarios pueden seleccionar diferentes niveles de consistencia dependiendo de los requisitos que tenga cada aplicación.
- **Alto rendimiento:** Cassandra se optimizó para operaciones de escritura y lectura lo que es ideal para aplicaciones que necesitan procesar grandes cantidades de datos en tiempo real.

Por otra parte, si hablamos de la arquitectura de Cassandra, cabe destacar que se organiza en varios componentes clave:

- **Nodos:** Cada nodo es responsable de gestionar y almacenar información. No hay una jerarquía fija entre nodos, por los que todos los nodos tendrán funciones equivalentes.
- **Anillo:** Los nodos se repartirán en una estructura en forma de anillo en donde cada uno administrara claves.
- **Replicación:** Los datos son replicados en diferentes nodos para asegurar su disponibilidad en caso de que se produzcan errores.
- **Particiones:** Los datos son divididos en particiones que serán repartidas entre los diferentes nodos del anillo.

Como dijimos anteriormente, el modelado de datos en Cassandra se centra en diseñar esquemas los cuales maximizan la eficiencia y escalabilidad. Lo cual contrasta con las bases de datos tradicionales, que utilizan un modelo basado en tablas que tienen filas y columnas, optimizado para repartir los datos de forma eficiente y rápida.

En resumen, a pesar de que Cassandra es una solución adecuada en varios posibles escenarios, no es la mejor opción si hablamos de construir un Data Warehouse tradicional, ya que es vital determinar desde el principio los casos de uso y que tipo de consulta

realizaremos, para poder diseñar un esquema de bases de datos que sea coherente u poder obtener su máximo rendimiento a la hora de trabajar con grandes volúmenes de datos [29].

4.7.4.3 MongoDB

MongoDB nació en el año 2007, fue fundado por la empresa 10gen, la cual se conoce a día de hoy como MongoDBInc, además en 2007 era parte de la playaforma PaaS(Platform as a Service). No obstante, en 2009, se modificó el enfoque del proyecto para convertirse en una base de datos independiente. Desde entonces se ha popularizado y es usado por grandes empresas por su capacidad de trabajar con grandes cantidades de datos de forma eficiente.

Cabe destacar que MongoDB es una base de datos NoSQL orientada a documentos que se utilizada en aplicaciones modernas debido a su flexibilidad y capacidad de escalabilidad. A diferencia de las bases de datos relacionales tradicionales, MongoDB almacena datos en documentos BSON (Binary JSON), lo que permite una representación de datos más jerárquica y compleja.

Las bases de datos NoSQL se introdujeron como alternativa a las bases de datos SQL relacionales tradicionales para poder almacenar y procesar grandes cantidades de datos de manera flexible y escalable. Las bases de datos NoSQL se caracterizan por su capacidad de almacenar datos sin necesidad de un esquema fijo, lo que es adecuado para aplicaciones con diferentes necesidades y datos no estructurados. Hay muchos tipos de bases de datos NoSQL, incluidas las basadas en documentos, las de clave-valor, las que se basan en grafos y también las orientadas a columnas [30].

A continuación, se muestra un ejemplo de los tipos de documentos que MongoDB puede almacenar en una colección:

```
{
    "title": "Troy",
    "director": " Wolfgang Petersen",
    "year": 2004,
```

```
    "genres": ["Action", "Adventure"],
    "cast": [
        {"name": "Brad Pitt", "role": "Aquiles"},
        {"name": "Eric Bana", "role": "Héctor"},
        {"name": "Diane Kruger", "role": "Helena"}
    ],
    "ratings": 7.3
}
```

El ejemplo anterior contiene información sobre la película Troya, incluyendo su título, director, año de lanzamiento, géneros, actores principales y puntuación.

En conclusión, MongoDB no es la primera opción para un Data Warehouse tradicional, ya que estos sistemas por lo general necesitan un esquema estructurado y una fuerte consistencia transaccional que las bases de datos relacionales proporcionan mejor. Sin embargo, MongoDB puede ser útil en escenarios donde se necesita flexibilidad en el esquema y se manejan grandes volúmenes de datos no estructurados o semiestructurados. Su capacidad de escalabilidad horizontal y alta disponibilidad también puede ser una ventaja en algunos contextos de Data Warehousing moderno.

4.7.4.4 HBase

Apache HBase es una base de datos no relacional la cual surgió en 2007 dentro del ecosistema de Hadoop. Esta fue creada para gestionar grandes volúmenes de datos estructurados a través de servidores o clústeres, lo que posibilita el acceso a los datos de manera ágil tanto a la hora de lectura, como escritura, en tiempo real. Originalmente se desarrolló como un proyecto de código abierto. Además, HBase usa un sistema de archivos de Hadoop el cual se conoce como HDFS.

Cabe destacar, entre las ventajas que ofrece HBase, su capacidad para escalar, lo que lo convierte en una herramienta ideal para aplicaciones que requieren trabajar con grandes volúmenes de daros, como dijimos anteriormente. También, su diseño atractivo posibilita el acceso inmediato a la información, lo que nos permite realizar consultas de manera ágil.

Es importante mencionar que se puede integrar con HDFS para que sea más robusto frente a posibles errores, es decir que el sistema podrá continuar activo incluso si se produce algún fallo.

Por último, el modelo columnar que datos que usa permite ser más eficiente en cuanto a realizar operaciones de lectura y escritura, esto nos permitirá optimizar el rendimiento del sistema [31].

4.7.4.5 Cassandra vs MongoDB vs HBase

A continuación, vamos a comparar tres de las bases de datos NoSQL más populares: Cassandra, MongoDB y HBase. Analizaremos sus similitudes y diferencias, para conocer cuál es más adecuada teniendo en cuenta nuestras necesidades.

Entre las similitudes encontramos:

- **Escalabilidad Horizontal**: Todas están diseñadas para escalar horizontalmente, esto quiere decir que tendremos la posibilidad de trabajar con mayores volúmenes de datos distribuidos. Esto es esencial para aplicaciones que requieren la gestión de grandes cantidades de datos.
- **Alta Disponibilidad y Tolerancia a Fallos**: Cada una de ellas están diseñadas para ser altamente disponibles y tolerantes a fallos, utilizando diferentes mecanismos para garantizar la disponibilidad de los datos incluso en caso de fallos del sistema.
- **NoSQL**: Las tres bases de datos son NoSQL, lo que significa que no utilizan el modelo relacional tradicional. Esto permite manejar datos no estructurados y semiestructurados de manera eficiente.

Por otra parte, también encontramos importantes diferencias:

Si hablamos de consistencia y disponibilidad, Cassandra se enfoca en la consistencia eventual, priorizando la disponibilidad y la partición tolerante. MongoDB ofrece configuraciones flexibles que permiten equilibrar consistencia y disponibilidad según nuestras necesidades. Por último, Hbase se centra en proporcionar una consistencia fuerte y alta disponibilidad dentro del ecosistema Hadoop.

Por otra parte, para los casos de usos, Cassandra es ideal para aplicaciones que requieren alta disponibilidad y velocidad de escritura, como aplicaciones de IoT y redes sociales. MongoDB se adapta bien a aplicaciones que necesitan una estructura de datos flexible y consultas complejas, como aplicaciones web y móviles. Hbase es utilizado en aplicaciones que requieren procesamiento en tiempo real y se benefician de la integración con el ecosistema Hadoop, como análisis en tiempo real y sistemas de recomendación.

En conclusión, para construir un Data Warehouse, Cassandra y HBase suelen ser más adecuadas debido a su capacidad para manejar grandes volúmenes de datos distribuidos y su enfoque en la escalabilidad horizontal. Ambas bases de datos pueden gestionar eficientemente operaciones de escritura y lectura a gran escala, lo cual es crucial para un Data Warehouse. Por otra parte, MongoDB ofrece flexibilidad en la estructura de datos y es ideal para aplicaciones que requieren consultas complejas y una rápida adaptación a cambios en los datos. Sin embargo, para un Data Warehouse que prioriza operaciones de lectura optimizadas y consistencia en los datos, Cassandra y HBase son opciones más robustas [32].

4.7.5 Herramientas de gestión de clúster

4.7.5.1 Hadoop

Hadoop es una tecnología de código abierto implementado por Douglas Cutting, fue diseñado para almacenar y procesar grandes volúmenes de datos. Esta herramienta permite a las empresas trabajar con los datos de manera eficiente [33].

El ecosistema de Hadoop incluye una variedad de componentes clave:

- **Hadoop Distributed File System (HDFS):** se conoce HDFS como el sistema de archivos distribuido de Hadoop, creado para gestionar grandes cantidades de datos en diferentes dispositivos. HDFS permite almacenar ficheros y es tolerante a errores, lo que nos garantiza la disponibilidad y fiabilidad de los datos.
- **MapReduce:** es el modelo de programación de Hadoop en el cual se dividen las tareas en partes más pequeñas las cuales se procesarán simultáneamente, lo que permite trabajar con los datos de manera más eficiente.
- **YARN (Yet Another Resource Negotiator):** YARN lo podemos definir como el sistema operativo de Hadoop, el cual se encarga de administrar los recursos del

clúster y programar tareas. Facilita que diversas aplicaciones aprovechen de manera eficiente los recursos. Cabe destacar que este proyecto no usaremos clúster.

- **Hadoop Common:** Componente que nos ofrece librerías esenciales para los demás módulos de Hadoop. Contiene las principales herramientas y las interfaces de programación (APIs) requeridas para el funcionamiento del sistema.

En la siguiente imagen podemos observar el ecosistema de Hadoop en el cual aparecen algunas herramientas que ya hemos mencionado y otras sobre las cuales hablaremos más adelante.

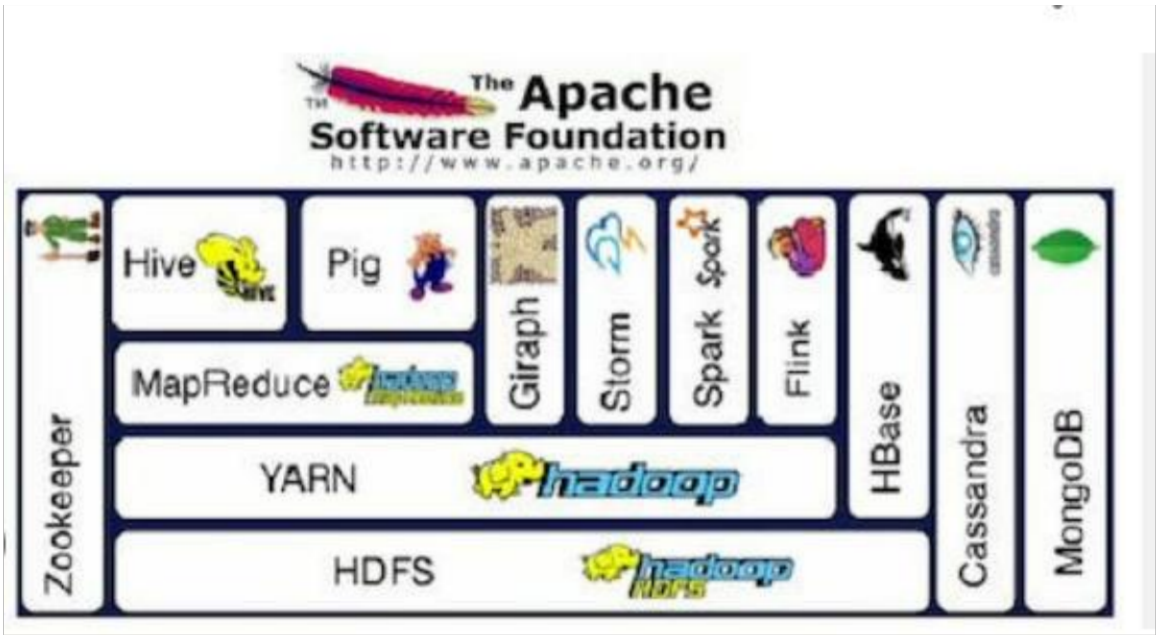


Figura 17. Ecosistema de Hadoop - apache.org

También es importante destacar que con el paso de los años en el ámbito del Big Data han ido apareciendo nuevas herramientas complementarias a Hadoop. Un ejemplo podría ser Spark, la cual se explicará más adelante. Estas herramientas, se centran en procesar datos de manera distribuida, lo que permite aprovechar técnicas como integrar datos en la nube o procesamiento en memoria. A diferencia de Hadoop, que está formado por otras herramientas como HDFS, MapReduce, como se dijo anteriormente.

Algunas de las ventajas que tienen estas nuevas herramientas es que tienen un procesamiento más rápido de la información, ya que muchas herramientas usan procesamiento en memoria lo que permite realizar consultas en un menor tiempo. También, tienen una gestión simplificada, pues al estar diseñadas para entornos en la nube hacen más fácil la administración de datos.

Por otra parte, los inconvenientes que tienen son la dependencia de proveedores en la nube, ya que estas herramientas suelen estar asignadas a un proveedor especifico, esto puede provocar dependencia y se poco flexible. Además, los costos son mucho más altos, aunque son más veloces y fáciles de usar, el uso de servicios en la nube, pueden provocar mayores costos en comparación con Hadoop.

En resumen, mientras Hadoop es una opción fiable para el procesamiento distribuido de datos, las otras herramientas pueden ser más adecuadas si disponemos de poco tiempo y priorizamos la velocidad de procesamiento.

4.7.5.1.1 HDFS

HDFS como dijimos anteriormente, es un sistema que permite gestionar grandes cantidades de datos. Este se inspiró en el sistema de archivos de Google, lo que le permite ser eficiente. Algunas sus principales características son:

- Alta tolerancia a fallos, ya que se realizan réplicas de datos en varios nodos, lo cual asegura su disponibilidad
- Manejo de grandes bloques de datos, lo que optimiza el rendimiento
- Arquitectura distribuida que permite un clúster de servidores, y en consecuencia permite un almacenamiento escalable y de alta disponibilidad.

Al cargar un archivo, HDFS lo divide en bloques, estos se distribuyen en diferentes nodos. El NameNode conserva la ubicación de estos bloques, de manera que los usuarios podrán acceder de manera eficiente a la información almacenada. Cabe destacar que NameNode y DataNode son programas que se ejecutan en equipos simples, por lo general con un sistema operativo GNU/Linux [34].

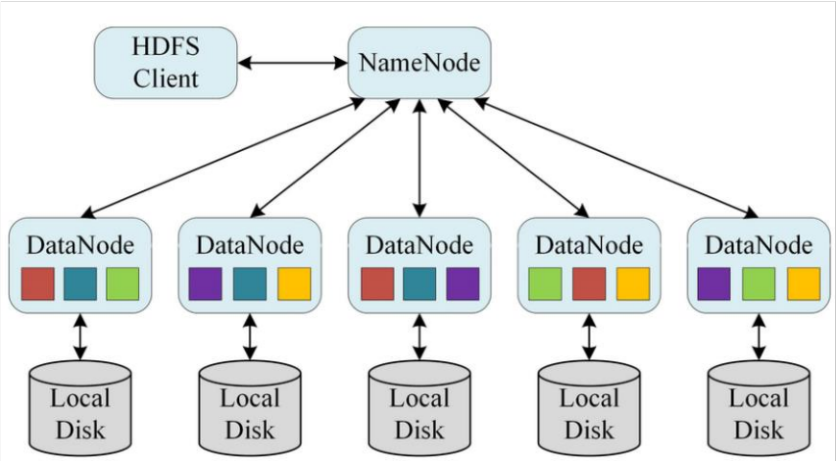


Figura 18. Ecosistema HDFS - medium.com

4.7.5.1.2 MapReduce

Hadoop MapReduce es un modelo de programación diseñado para procesar y generar grandes volúmenes de datos de manera paralela y distribuida a partir de un clúster. En sus inicios, fue desarrollado por Google y ha sido implementada en el ecosistema Hadoop, convirtiéndose en una herramienta fundamental para procesar los datos de manera eficiente. En este modelo encontramos dos etapas. Map y Reduce.

En Map los datos de entrada son descompuestos en pares clave-valor. Además, cada tarea se divide en unidades más pequeñas las cuales pueden ejecutarse en paralelo, lo que mejora la eficiencia.

Por otro lado, encontramos la fase Reduce, cuando ya ha terminado el mapeo, los datos resultantes se agrupan para ser combinados. En esta etapa se procesan y consolidan los resultados en grupos más pequeños, generando una salida final [35].

Un ejemplo de MapReduce es el conteo de palabras en un conjunto de documentos de grandes. En la fase Map, cada documento se analiza y cada palabra encontrada se convierte en un par clave-valor, con la palabra como clave y el número 1 como valor. Después, en la fase "Reduce", los pares clave-valor se agrupan por palabra y se suman los valores para obtener el conteo total de cada palabra en todos los documentos.

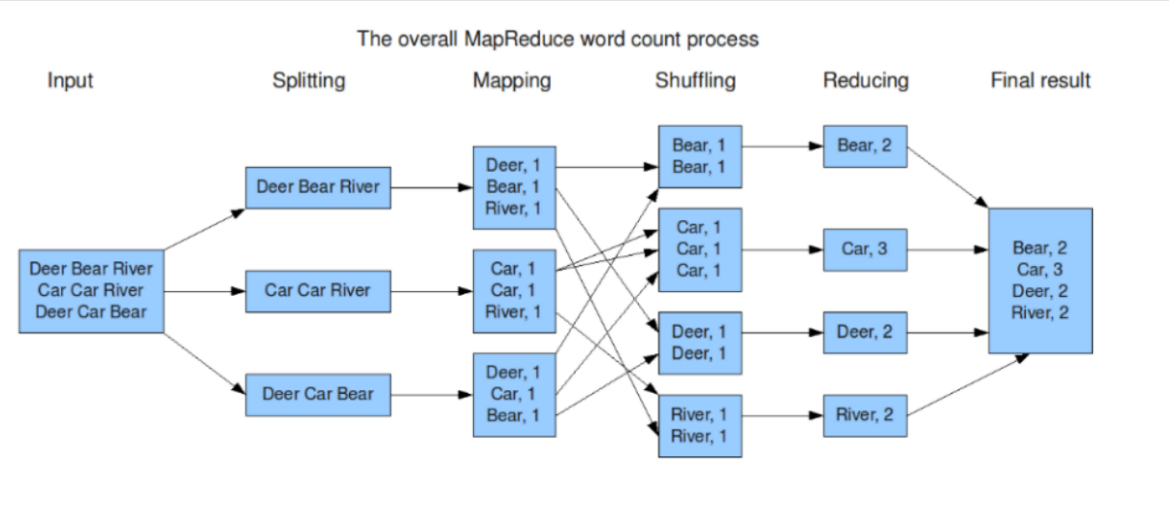


Figura 19. Ejemplo Map Reduce - codesyntax.com

4.7.5.1.3 YARN

YARN (Yet Another Resource Negotiator), es un elemento clave en el ecosistema de Hadoop, ya que su principal función es hacer más óptima la gestión de recursos y coordinar diferentes tareas, superando limitaciones que podían tener las herramientas comentadas anteriormente.

Además, YARN automatiza la asignación de recursos y planificación de tareas, lo que posibilita hacer más eficiente el espacio disponible. Por ejemplo, ResourceManager reparte los recursos entre las aplicaciones que se ejecutan en un clúster, lo que permite que haya distintas herramientas de procesamiento como MapReduce que opera simultáneamente en un único entorno.

También, YARN se encarga de la supervisión de procesos, detectando posibles fallos y corrigiéndolos automáticamente si es necesario. Esto nos asegura una gran disponibilidad y ejecutar múltiples tareas a la vez [36].

Dentro de la arquitectura YARN encontramos diferentes componentes:

- **ResourceManager:** Es el núcleo central en la gestión, el cual asigna recursos del clúster.
- **NodeManager**: Es ejecutado en cada nodo del clúster y monitoriza el uso de recursos como CPU, memoria, etc, en cada contenedor. Posteriormente enviando esta información al ResourceManager.

- **ApplicationMaster:** Se encarga de la ejecución y supervisar una aplicación específica. Además, gestiona los recursos necesarios.
- **Contenedor:** Es una unidad de recursos físicos en la cual se ejecutan las aplicaciones, esto facilita el control de los recursos.

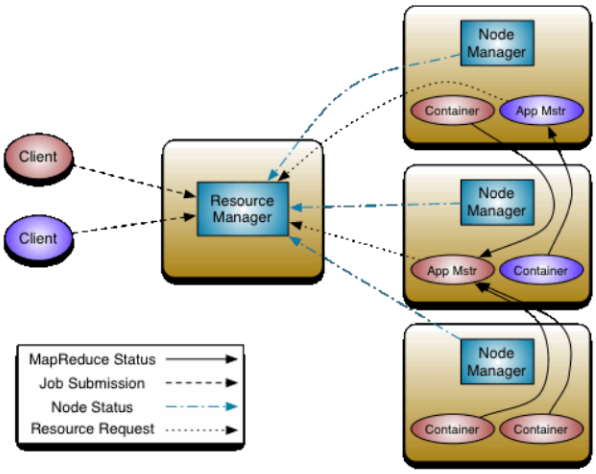


Figura 20. Ecosistema YARN - hadoop.apache.org

4.7.5.2 Apache Ambari

Ambari es una herramienta la cual nos permite gestionar, monitorizar y configurar de manera sencilla clústeres en el ámbito del Big Data que se basan en el ecosistema de Hadoop. Esta herramienta nos proporciona una interfaz muy intuitiva, que nos permite en todo momento comprobar el estado del clúster, realizar tareas de mantenimiento y poder añadir nuevos servicios o nodos.

Además, Ambari destaca porque nos permite automatizar tareas que pueden resultar complejas como la instalación de algunas herramientas que ya comentamos anteriormente como HDFS, YARN, Spark, etc. Esto nos ayuda a no gastar tanto tiempo comparado si lo hacemos de forma manual. También Ambari tiene herramientas que nos permite monitorear el rendimiento del clúster en tiempo real, lo que nos permite detectar posibles fallos en nodos.

Por otro lado, esta herramienta se puede integrar con sistemas de autenticación como Kerberos, lo que nos puede resultar útil si priorizamos aspectos como la seguridad [37].

4.7.6 Herramientas ETL

4.7.6.1 Pentaho Data Integration

Pentaho es una herramienta ETL de código abierto, la cual nos permite diseñar y ejecutar flujos de trabajo de extracción, transformación y carga de datos (ETL), para ello la herramienta nos proporciona una interfaz gráfica conocida como Spoon, en la cual podemos colocar diferentes componentes, conectarlos a diferentes fuentes sin la necesidad de que tengamos que programar.

Pentaho nos ofrece una amplia variedad de conectores, nos permite cargar información en diferentes bases de datos relacionales y no relacionales, archivos como CSV o XML, etc. Esto nos facilita mucho la integración de los datos.

Pentaho trabaja con un repositorio centralizado o con archivos locales, en donde podemos almacenar las transformaciones y los jobs. Así, podemos planear su ejecución de manera unificada.

Además, Pentaho destaca porque nos permite organizar los trabajos. Incluye soporte para algunas herramientas Big Data que comentamos anteriormente como Spark, Hive, etc. También Pentaho cuenta con pantallas en las que podamos ver cada job y transformación, podemos ver también los tiempos de ejecución.

En conclusión, Pentaho es una herramienta que nos puede ser útil para realizar los ETLS ya que su ofrece una interfaz intuitiva que, aunque no tengamos muchos conocimientos de programación nos permite realizar la función de cargar los datos transformados en una base de datos [38].

A continuación, vemos una imagen de un ejemplo de flujo ETL en donde podemos ver algunas de las funcionalidades que nos ofrece la herramienta.

Figura 21. Ejemplo flujo ETL - Elaboración propia

4.7.6.2 Pipelines de ETL

Otra herramienta interesante para realizar los procesos ETL son los pipelines de ETL. Un pipeline es un conjunto de pasos mediante el cual extraemos los datos de diferentes fuentes, los transformamos y posteriormente los podemos cargar en un almacén de datos como Hive.

En vez de diseñar las transformaciones de manera manual a través de una interfaz gráfica, como ocurría con Pentaho, como explicamos anteriormente, podemos usar Spark, que es el que utilizaré en este proyecto, para montar pipelines como código, ya sea con código Python o Scala [39].

4.7.7 Herramientas de Cubos OLAP

4.7.7.1 Microsoft SQL Server Management Studio (SSMS)

Una herramienta muy usada es Microsoft SQL Server Management Studio (SSMS) para gestionar y configurar instancias de SQL Server y las bases de datos asociadas. Esta herramienta nos ofrece una interfaz gráfica que nos ayuda a la creación de consultas, administración de bb.dd, etc. Además, que podemos generar scripts administrativos y elaborar informes avanzados, lo que nos simplifica muchas tareas.

SSMS es una herramienta importante para gestionar bases de datos, ya que tiene funciones necesarias para utilizar SQL Server. Se utiliza sobre todo por administradores y desarrolladores de bases de datos por su facilidad de uso. También se puede integrar con Azure SQL, lo que nos puede ayudar a gestionar bases de datos en la nube de manera sencilla [40].

Entre las características de SSMS destacamos:

- **Consultas SQL:** Nos permite escribir, ejecutar y depurar consultas SQL en un entorno de pruebas eficiente.

- **Explorador de objetos:** Nos hace más sencilla la navegación por los diferentes componentes de la base de datos como tablas, variables, procedimientos, etc.
- **Monitoreo del rendimiento:** Las herramientas nos ayudan a analizar y mejorar el rendimiento de las consultas y del propio servidor.
- **Compatibilidad con cubos OLAP:** Además de gestionar bases de datos relacionales, SSMS nos permite administrar cubos OLAP a través de SQL Server Analysis Services, esto es esencial para analizar grandes volúmenes de datos en entornos multidimensionales.

Por otra parte, si nos referimos a la arquitectura, SSMS es como un puente entre los usuarios, los administradores, y SQL Server. Cabe destacar que funciona bajo un modelo cliente-servidor. Es decir, SSMS actúa como el cliente, lo que nos permite enviar solicitudes de administración, mientras que SQL Server se encarga de almacenar la información y procesar las consultas. Además, con la integración con SQL Server Analysis Services, podemos trabajar con cubos OLAP para realizar análisis multidimensionales y poder generar informes complejos.

En resumen, SSMS es una herramienta la cual nos permite gestionar tanto bases de datos transaccionales (OLTP) como también multidimensional (OLAP), siendo muy útil si usamos SQL Server.

4.7.7.2 Kylin

Apache Kylin es un motor analítico de código abierto que nos permite ejecutar consultas SQL y realizar análisis multidimensional OLAP en grandes volúmenes de datos sobre Hadoop.

Inicialmente fue desarrollado por eBay, donde en 2013 sus ingenieros pensaron en ofrecer análisis en tiempo real a sus usuarios sobre gran cantidad de datos. Como las soluciones tradicionales de OLAP no se adaptaban a estos volúmenes, surgió Kylin, mediante la agregación de cubos OLAP, lo que nos permite realizar consultas muy rápidas. Posteriormente en 2015, eBay donó el proyecto a Apache Software Foundation y a partir de ahí han estado en continua evolución [41].

En cuanto a la arquitectura de Kylin, se centra en la construcción de cubos OLAP, que se pueden consultar rápidamente por herramientas de Business Intelligence. Entre sus componentes podemos destacar:

- **Cubos OLAP**: Kylin consolida los datos en cubos multidimensionales los cuales son almacenados en sistemas como HDFS, esto nos permite que las consultas se realizan de manera muy rápida, ya que parte del procesamiento se realiza previamente.
- **Motor de Consulta**: Transforma consultas SQL en operaciones sobre cubos, esto nos permite obtener respuestas en tiempo real.
- **Motor de Ingesta de Datos**: Kylin integra información de diversas fuentes como Hadoop, Hive, bases de datos relacionales para poder construir los cubos.
- **Interfaz de Usuario**: Hay una interfaz web la cual nos ayuda a la creación y gestión de cubos, además de ejecutar consultas de manera sencilla.

En la figura podemos ver la estructura de Apache Kylin.

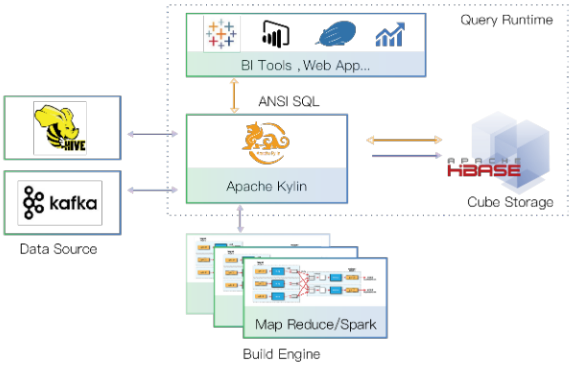


Figura 22. Estructura Apache Kylin - todobi.com

Kylin nos ofrece ventajas. Por una parte, como dijimos, su capacidad para realizar consultas OLAP a una velocidad elevada, hace que sea una herramienta muy valiosa para análisis en tiempo real. Además, tiene una alta escalabilidad que nos permite manejar datos distribuidos en Hadoop. Por otra parte, nos permite integrarla fácilmente con otras herramientas BI, lo que nos puede facilitar su incorporación.

4.7.7.3 Kylin vs Microsoft SQL Server Management Studio

Por lo que se dijo anteriormente, es útil comparar estas dos herramientas, teniendo en cuenta diferentes aspectos, para saber si es mejor usar una u otra. Por una parte, Kylin es una herramienta de análisis OLAP, la cual nos permite ejecutar consultas SQL de manera rápida sobre datos que tenemos almacenados en Hadoop. Se centra en acelerar el procesamiento de datos.

Por otro lado, SSMS nos ofrece una interfaz gráfica, es más intuitivo para interactuar con SQL Server, haciéndonos mucho más fácil la administración de bases de datos. También podemos crear consultas, generar informes, entre otras funcionalidades, de manera más intuitiva.

En cuanto a rendimiento, Kylin destaca al ejecutar análisis complejos, ya que usa cubos ya calculados para evitar procesar datos en tiempo real, esto es muy útil si trabajamos con grandes volúmenes de datos. Por otro lado, SSMS es más óptimo para manejar bases de datos tradicionales, y a pesar de que es muy eficiente en ese ámbito, no ofrece el mismo nivel de rendimiento si trabajamos con grandes volúmenes de datos.

En resumen, si queremos construir un Data Warehouse, usar SSMS puede ser la mejor opción por ser más fácil y eficiente a la hora de trabajar con bases de datos. Sin embargo, su tenemos que trabajar con grandes cantidades de datos, sería mejor usar Kylin por su alta velocidad de procesamiento.

4.7.8 Herramientas de desarrollo

4.7.8.1 Zeppelin Notebook

Apache Zeppelin es una herramienta de código abierto la cual nos permite crear cuadernos interactivos que están diseñados para analizar datos en un entorno Big Data. Con esta herramienta podemos escribir scripts en diferentes lenguajes de programación como Scala, Phyton, SQL, etc. Además, podemos realizar consultas, visualizar gráficos y documentar nuestros análisis en un entorno colaborativo, lo que nos facilita mucho el trabajo en entornos Big Data.

Una de las principales ventajas que tiene Zeppelin, es que se integra de forma sencilla con otras Apache Spark, lo que facilita la creación y ejecución de procesos ETL y posterior análisis sobre cantidades de datos masivas. Por otro lado, Zeppelin también se puede conectar a otras herramientas Big Data como Hive, Cassandra, Hadoop, etc. Lo que nos

permite realizar consultas con los datos que tengamos en la base de datos que estemos utilizando.

Utilizar cuadernos interactivos como Zeppelin no solo nos ayuda en la colaboración de un proyecto, sino que también se acelera el proceso de validación de datos.

Como se dijo anteriormente, Zeppelin es muy útil si se trabaja de forma colaborativa, ya que permite compartir código, análisis, gráficos, y resultados en tiempo real. También, dispone de un soporte para visualizaciones dinámicas y personalizadas para ver los datos de manera gráfica, en el cual aparecen diferentes gráficos, como barras, líneas, etc. Esto ayuda mejora considerablemente la presentación de datos y nos permite acelerar los procesos de exploración y documentación en entornos colaborativos.

Zeppelin también permite ejecutar el código en diferentes Jobs, en los cuales podemos tener en cada uno un lenguaje de programación diferente, esto nos a ayuda mucho a la hora de deputar errores, realizar consultas, etc.

Zeppelin se usa mucho en entornos empresariales por su flexibilidad y su capacidad de conectarse a diferentes fuentes de datos [42].

4.7.8.2 Jupyter Notebook

Similar a Zeppelin, Jupyter es una herramienta muy popular a la hora de analizar datos. En este caso, se combina el código con visualizaciones y texto de manera uniforme en un único documento, lo que nos facilita mucho el proceso de visualización de datos [43].

En el caso de que tengamos muchas líneas de código, Jupyter nos permite ejecutar código de manera fragmentada y poder visualizar el resultado de forma inmediata. También, nos permite integrar gráficos y descripciones en un documento.

Jupyter soporta varios lenguajes de programación, como Python, R, Scala o Julia entre otros, pero el que más se utiliza es Python. También, permite ejecutar el código en bloques, lo que ayuda a la hora de depurar errores e ir avanzando de manera progresiva a la hora de realizar los análisis.

Además, Jupyter cuenta con extensiones como JupyterLab, que se pueden añadir en terminales o paneles, permitiendo al usuario tener una experiencia similar a la que tendría si estuviera utilizando un IDE.

Jupyter además cuenta con una gran variedad de librerías como Matplotlib, Pandas, o NumPy, lo que facilita mucho la programación. Además, es compatible con otras herramientas de procesamiento como Spark, o bases de datos como Hive, mediante el uso de extensiones o kernels especiales para ello.

4.7.8.3 Databricks Notebooks

Los Databricks nos ofrecen un entorno colaborativo en la nube el cual está integrado con Spark, lo que nos permite construir pipelines de datos y trabajar de manera colaborativa. Gracias a los Databricks, como en otros notebooks se puede combinar código, visualizaciones y documentos en un único espacio de trabajo [44].

Además, los Databricks automatizan la gestión de clusters, los cuales pueden configurar políticas de escalado, snapshots y control de versiones de notebooks.

Los notebooks de Databricks soporta Delta Lake para almacenar de forma transaccional,y utiliza MLflow para realizar un seguimiento completo del ciclo de vida de los modelos Machine Learning. También ofrecen conectores nativos a servicios en la nube como AWS S3 y Azure Data Lake, como también permite la integración con Kafka y bases de datos NoSQL.

Desde los notebooks de Databricks, también se pueden programar ejecuciones como jobs, lo que permite automatizar flujos de trabajo directamente desde el entorno.

4.7.8.4 Visual Studio

Por otra parte, Visual Studio es un entorno de desarrollo creado por Microsoft, el cual se utiliza principalmente para implementar aplicaciones y gestionar proyectos software. Visual Studio cuenta con una gran variedad de lenguajes de programación, como C#, Python, etc. Además, permite la integración de bases de datos y la configuración de repositorios de código de manera sencilla, lo que nos es muy útil a la hora de trabajar en entornos colaborativos [45].

Además, Visual Studio incorpora herramientas para depuración de código como breakpoints, condicionales, inspección de variables en tiempo real. También se pueden añadir numerosas

extensiones como asistentes de generación de código y extensiones para Docker, Azure DevOps, etc.

En cuanto a la parte de "Data Science", incluye un visor de DataFrames, tiene capacidad para realizar análisis de datos de forma interactiva. Esto permite trabajar con notebooks directamente dentro del IDE, combinar documentación y visualizaciones con el flujo de desarrollo habitual y realizar depuración usando las herramientas mencionadas anteriormente.

Además, Visual Studio ha incorporado notebooks de Jupyter, los cuales se pueden usar a través de una extensión llamada "Jupyter". Esto realizar las mismas tareas que en Jupyter, pero dentro de Visual Studio, aprovechando así las diferentes extensiones que dispone, la fácil integración con herramientas de control de versiones como Git y demás herramientas que dispone Visual Studio, pero con la velocidad que tiene un notebook.

4.7.8.5 Notebooks vs IDEs

La elección entre cuadernos interactivos y un IDE debe basarse en las necesidades técnicas del proyecto y su flujo de trabajo.

En primer lugar, los notebooks, como Zeppelin, Jupyter o Databricks, permiten ejecutar el código por bloques de forma independiente. También permiten visualizar los resultados al instante. Este modelo de ejecución por bloques acelera el análisis, facilita la comprobación de hipótesis y reduce considerablemente los tiempos de espera al no tener que compilar ni poner en marcha una aplicación para cada prueba. Por otro lado, los IDEs como Visual Studio ejecutan el programa en su totalidad o en módulos empaquetados, lo que resulta más adecuado para flujos de desarrollo, compilación, pruebas unitarias, etc.

El depurado de los notebooks es limitado, generalmente se usa la impresión de variables o el uso de herramientas de depuración como JupyterLab, que no alcanzan ni de lejos la profundidad de un IDE. Visual Studio, por su parte cuenta con depuradores avanzados los cuales incluyen breakpoints condicionales, análisis de rendimiento de CPU y memoria, visualización gráfica de llamadas, etc. Las capacidades de un IDE son fundamentales cuando la aplicación crece en complejidad y es necesario identificar cuellos de botella en el sistema.

Por último, los notebooks incorporan conectores directos a herramientas de Big Data como Spark, Hive y pueden ejecutarse sobre clústeres distribuidos, esto agiliza el procesamiento

de grandes volúmenes de datos. Mientras que los IDEs requieren de extensiones o de alguna configuración por parte del usuario para conectarse con estas tecnologías, por otra parte, se integran mejor con herramientas de despliegue.

4.7.9 Virtualización y contenedores

4.7.9.1 Máquinas virtuales

Las máquinas virtuales son entornos de software los cuales simulan un ordenador completo, incluyendo su propio sistema operativo, kernel, almacenamiento, memoria, CPU, y red. Las máquinas virtuales cuentan con un monitor conocido como hipervisor, el cual permite ejecutar múltiples sistemas operativos en un único equipo. Este tipo de virtualización permite un aislamiento completo entre los diferentes sistemas, lo cual es de gran utilidad para realizar pruebas, crear entornos de desarrollo independientes, ejecutar software que solo funciona en un sistema operativo específico, etc.

Cada sistema operativo de la máquina virtual se puede configurar asignándole una cantidad de recursos (memoria RAM, almacenamiento, núcleos de CPU), y además puede funcionar de manera independiente del sistema anfitrión [46].

Las máquinas virtuales más populares son VirtualBox y VMware nos permite realizar acciones como:

- Crear instantáneas o snapshots del estado actual de la máquina para tener la posibilidad de retroceder a ese estado de la máquina en caso de que se produzca algún fallo en el futuro.
- Simular redes en las máquinas virtuales
- Definir políticas de disponibilidad, tolerancia a fallos y recuperación de forma automática ante fallos. Estas son algunas de las características que hacen que las máquinas virtuales sean utilizadas en entornos colaborativos y en pruebas de software que requieren un gran aislamiento.

4.7.9.2 Docker

Docker es una herramienta de código abierto que se encarga de automatizar el despliegue de aplicaciones dentro de contenedores software. A diferencia de las máquinas virtuales, los

contenedores no necesitan virtualizar todo el sistema operativo, esto debido a que comparten el kernel del sistema anfitrión y son ejecutados como procesos aislados, esto los hace más ligeros y más rápidos de iniciar.

Cada contenedor se encarga de empaquetar la aplicación con todas sus librerías, dependencias, configuraciones y demás herramientas para su ejecución, garantizando de esta manera que se comporte igual que cualquier entorno. Esto permite conseguir una gran portabilidad y consistencia en entornos de desarrollo [47].

Docker tiene una arquitectura cliente-servidor, en la cual el cliente va ejecutando comandos y el servidor llamado Docker Daemon se encarga de gestionar los contenedores.

Docker destaca por:

- Sistema de imágenes por capas, en donde cada instrucción genera una capa reutilizable, esto reduce espacio y acelera la construcción de imágenes.
- Repositorio centralizado, desde el cual se pueden descargar miles de imágenes ya configurados.
- Docker Compose, permite definir múltiples contenedores en un único archivo para levantarlos como un sistema completo.
- Volúmenes de datos persistentes, redes que han sido definidas por el usuario y compatibilidad con drivers externos.
- Integración con herramientas como Kubernetes y Swarm, para escalar contenedores de forma automática en función de la carga.

4.7.10 Herramientas de visualización y BI

4.7.10.1 Power BI

Power BI es una herramienta de Business Intelligence desarrollada por Microsoft, la cual nos permite visualizar datos de manera interactiva y realizar análisis en tiempo real. Su principal ventaja es la posibilidad de conectarse a diferentes fuentes de datos, transformarlos y mostrarlos en forma de gráficos, informes, etc. De manera entendible y dinámica [48].

En Power Bi podemos destacar los siguientes componentes:

- **Power BI Desktop:** es una aplicación la cual instalamos en nuestros equipos para crear informes y poder diseñar dashboards. Aunque se utiliza de forma local para el desarrollo y la creación de contenidos, estos informes están destinados a publicarse en la nube para facilitar su acceso y distribución.
- **Power BI Service:** Es la plataforma en la nube de Power BI, donde se van alojando informes y dashboards creados en Power BI Desktop. Este elemento, nos permite mantener la información actualizada y accesible.
- **Power BI Mobile:** Es la aplicación que permite visualizar e interactuar con los informes generados en nuestros dispositivos móviles, esto nos asegura que la información esté disponible en cualquier lugar y momento.

Cabe destacar, que PowerBI no solo nos permite visualizar y analizar datos, además nos ofrece potentes herramientas de transformación y modelado de datos. Como Power Querry y Power Pivot, lo que nos permite que la información sea más limpia.

Otra característica importante es que se puede integrar con otros productos de Microsoft como por ejemplo Teams, lo que ayuda mucho a la hora de trabajar de forma colaborativa.

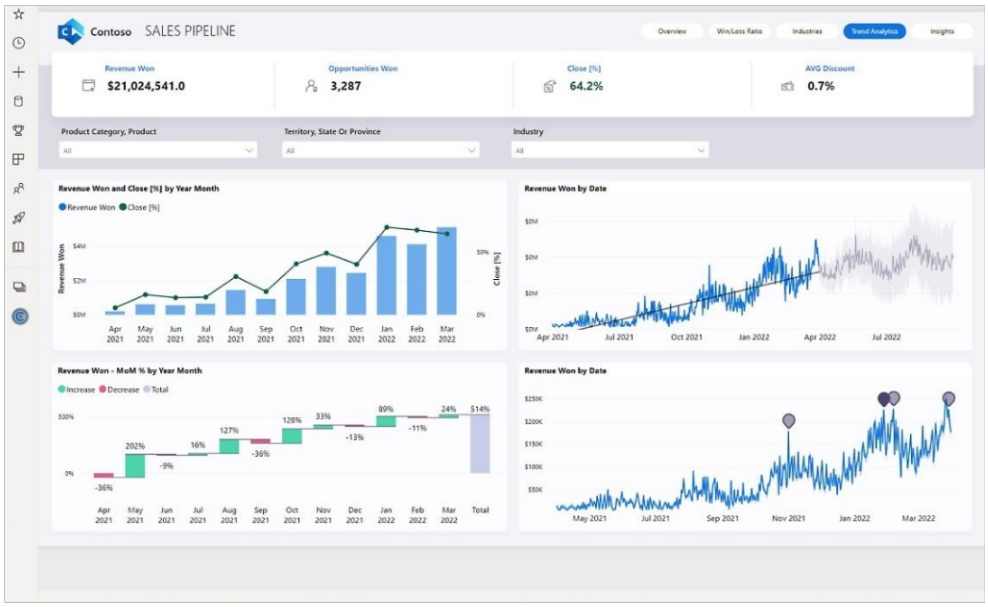


Figura 23. Ejemplo Dashboard Power BI - microsoft.com

4.7.10.2 Tableau

Tableau es una herramienta muy reconocida a la hora de visualizar los datos. Esta herramienta nos permite transformar grandes volúmenes de datos en gráficos interactivos o dashboards lo que nos puede ser de gran utilidad a la hora de tomar decisiones. Puede conectarse a diferentes fuentes de datos como bases de datos, archivos en la nube, etc. Esto nos permite realizar análisis de forma rápida y eficaz.

De Tableau Podemos destacar su interfaz intuitiva, la cual nos permite crear visualizaciones de manera sencilla, sin necesidad de tener conocimientos avanzados en programación. Esto nos facilita mucho el trabajo. Además, Tableau se puede usar en modo colaborativo ya que permite compartir dashboards y actualizar la información en tiempo real.

Por otro lado, podemos destacar que Tableau nos proporciona muchos recursos, como tutoriales y soluciones que nos ayudan a optimizar al máximo esta herramienta.

En resumen, Tableau es una buena opción para mejorar y optimizar nuestros análisis de datos además de que nos ayuda en la toma de decisiones si no tenemos muchos conocimientos de programación [49].

4.7.10.3 Qlick Sense

Qlik Sense es una herramienta moderna que como las anteriores nos permite explorar y visualizar datos de forma interactiva. Click Sense trabaja con grandes volúmenes de datos, cuenta con una interfaz fácil de usar y como Tableau permite crear dashboards sin tener unos conocimientos muy avanzados.

Qlick Sense se adapta a diferentes dispositivos, lo que nos permite trabajar desde diferentes equipos. La herramienta integra una amplia variedad de fuentes de datos [50].

5. Diseño de la arquitectura

5.1 Esquema de la arquitectura

En el apartado anterior vimos el estado del arte de un sistema de arquitectura BI tradicional. En este proyecto, el flujo de datos que se llevará a cabo lo podemos ver en la siguiente imagen.

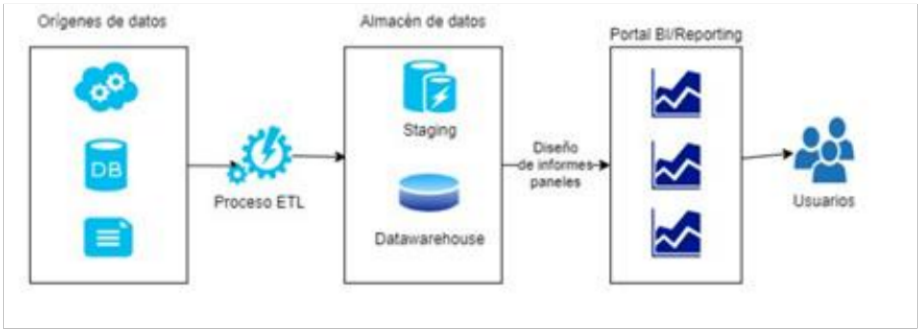


Figura 24. Diagrama de arquitectura BI - blogvisionarios.com

Como podemos ver, en primer lugar, cargaremos los datos. Para ello utilizaremos diferentes ficheros tsv públicos de IMDB que contiene información relevante sobre películas, series, actores, etc. Además de otros ficheros que hemos obtenido información mediante webscraping el cual nos ha llevado meses recopilar toda esa información. En segundo lugar, almacenamos los datos en nuestro Data Warehouse. Por último, visualizaremos los datos en un Dashboard.

5.2 Herramientas utilizadas

Todas las herramientas usadas son Open Source y nos permite escalar de una manera sencilla el sistema.

A la hora de procesar la herramienta utilizada es Apache Spark, la cual usamos para procesar los ficheros fuentes y adaptarlos a nuestro almacén de datos, este proceso ya lo mencionamos en puntos anteriores y se conoce como ETL. El motivo de utilizar Spark, es por su capacidad de procesar grandes volúmenes de datos en un tiempo pequeño.

Hadoop es una herramienta que además de permitirnos gestionar ficheros con HDFS. Además, permite añadir nuevos nodos y crear un clúster, esto permitiría agilizar la ejecución de los procesos, evitar posibles fallos en el sistema, cuellos de botella, etc.

Como notebook, se utiliza Apache Zeppelin, para implementar los ETLs y realizar operaciones con la base de datos. El lenguaje que he utilizado para realizar los ETLs es Scala ya que es el que viene por defecto con Spark y es más completa y se ejecuta más rápido. También ofrece un mejor rendimiento ya que con Python se puede producir sobrecarga. Además, Scala dispone de muchas librerías las cuales nos puede facilitar mucho a la hora de programar. El uso de Apache Zeppelin como notebook, es porqué nos permite visualizar los resultados de manera gráfica o en formato tabla, lo que ofrece una interfaz atractiva a la hora de visualizar los resultados.

Como almacén de datos, usaremos Hive que usa el lenguaje SQL. Además, utiliza HDFS como sistema de almacenamiento de ficheros. Esta herramienta es muy útil si queremos realizar consultas sobre el almacén de datos.

A continuación, a modo de ejemplo podemos observar la carga de un dataframe en una tabla de Hive y posteriormente su visualización en formato tabla mediante una consulta select, para ello se usará un fichero que he creado que contiene información de usuarios con su nombre y edad.

```
%spark

// Importar bibliotecas necesarias
import org.apache.spark.sql.SparkSession

// Crear el SparkSession con soporte para Hive
val spark = SparkSession.builder()
    .appName("Cargar CSV en DataFrame a Hive")
    .config("spark.sql.catalogImplementation", "hive") // Configuración para usar Hive
    .enableHiveSupport()
    .getOrCreate()

// Definir la ruta del fichero
val filePath = "file:///home/asa117/Descargas/usuarios.csv"

// Leer el archivo CSV y crear el DataFrame
val df = spark.read
    .option("header", "true")  // Si el CSV tiene encabezado
    .option("inferSchema", "true")  // Inferir el tipo de datos automáticamente
    .csv(filePath)

df.show()

// Escribir el DataFrame en una tabla Hive
df.write
    .mode("append")  // 'append' añadimos datos sobres los ya existentes si los hay
    .saveAsTable("ejemplo_db.ejemplo")  // tabla a la que vamos a cargar los datos
```

```
+---+------+----+
| id|nombre|edad|
+---+------+----+
|  1|   Ana|  25|
|  2|Carlos|  30|
|  3| Lucía|  28|
|  4|  Juan|  35|
+---+------+----+
```

Figura 25. Ejemplo carga dataframe usurios – Elaboración propia

```
%hive
use ejemplo_db;

SELECT * FROM ejemplo;
```

Query executed successfully. Affected rows : -1

settings

id	nombre	edad
1	Ana	25
2	Carlos	30
3	Lucía	28
4	Juan	35

Figura 26. Ejemplo select usuarios – Elaboración propia

Para la realización del cubo OLAP, se ha utilizado en primer lugar SQL Server Management Studio para cargar las diferentes tablas de Hive del almacén de datos, posteriormente para realizar el diseño del cubo y poder realizar consultas se ha utilizado Visual Studio en modo multidimensional y se creará posteriormente un proyecto de Analysis Services para diseñar y desplegar el cubo. El uso de un cubo OLAP con estas herramientas es para realizar consultas de manera más rápida que si utilizamos una BD.

Para la parte de visualización usaremos Power BI, debido a su capacidad para crear dashboards interactivos, además de que ofrece una amplia variedad de herramientas visuales. Power BI únicamente se puede usar en Windows, y algunas inconveniencias son la limitación de ciertas herramientas visuales y una interfaz gráfica no tan intuitiva, además de que el dashboard generado no se puede hacer público si no utilizamos la versión de pago.

6. Implementación del sistema

6.1 Fuentes de datos

En esta sección describiremos las fuentes de datos utilizadas, tanto aquellas públicas como las que obtuvimos información por webscraping, además de los ficheros resultantes que contienen la información que posteriormente extraeremos, limpiaremos y cargaremos en el almacén de datos.

6.1.1 IMDb – Non Commercial Datasets

En primer lugar, se obtuvieron datasets oficiales de IMDb, una de las bases de datos más grandes del mundo en la cual contiene información sobre millones de películas, series, actores, directores y mucha más información relacionada con el mundo audiovisual. Dichos datasets se pueden descargar en la sección "Non Commercial Datasets" [51].

Los datasets que encontramos en dicha sección son:

- title.akas.tsv: contiene información sobre los títulos de los contenidos en diferentes países.
- title.basics.tsv: contiene información principal de los contenidos, como el tipo, el nombre, etc.
- title.ratings.tsv: contiene la puntuación media y los votos de los contenidos.
- title.crew.tsv: contiene información de los directores y guionistas.
- title.principals.tsv: contiene información sobre los actores.
- name.basics.tsv: contiene información más detallada de las personas que participan en los contenidos, como su profesión y títulos en los que participan.
- title.episode.tsv: contiene información sobre los episodios de las series, como la temporada.

Dichos ficheros los podemos descargar de manera individual de la sección "Non Commercial Datasets", y estos se actualizan diariamente, lo que garantiza que los datos sean muy recientes y tenga una alta fiabilidad.

En los ficheros descargados tendremos información de millones de películas, series, cortometrajes, videojuegos y demás. De las cuales posteriormente se eliminará todo aquello que no sea una película o serie.

De los ficheros mencionados anteriormente, solo se utilizarán los que sean necesarios. Los ficheros que utilizaremos son: title.basics.tsv, title.ratings.tsv y name.basics.tsv.

6.1.1.1 title.basics.tsv

Este es el fichero que principal que se usará para realizar los análisis de los contenidos audiovisuales. Contine 1171070 filas, con contenidos de todo tipo como películas, series, cortometrajes, videojuegos y muchos más que como dijimos anteriormente, aquellos que no sean películas y series se eliminarán, los cuales abarcan desde 1974 hasta 2031 (incluyendo información de contenidos que aún no se han estrenado). Como se dijo anteriormente, este fichero se actualiza diariamente, por los que sus datos son muy recientes y fiables.

A continuación, podemos ver la tabla de metadatos de title.basics.tsv, en donde podemos observar detalladamente cada campo del fichero.

Fichero	Nombre columnas	Tipo datos	Descripción	Ejemplo
titles.basics.tsv	Tconst	String	id de la película o serie en IMDB	tt0332452
	titleType	String	tipo (película o serie)	Movie
	primaryTitle	String	título más popular en diferentes países	Troy
	originalTitle	String	título original	Troy
	isAdult	Boolean	indica si es para adultos (0->NO 1->SÍ)	0
	startYear	Int	año de estreno	2004
	endYear	Int	año de finalización	Null
	runtimeMinutes	Int	duración en minutos	163
	Genres	String	géneros de la película o serie	Adventure,Drama
	Languajes	String	idiomas del contenido audiovisual	English

Tabla 1. Metadatos title.basics.tsv

6.1.1.2 title.rating.tsv

Este fichero contiene 1454241 filas aquí encontramos la puntuación y el número de votos de todos los contenidos de IMDb. Este fichero se actualiza diariamente, lo que asegura que las puntuaciones y votaciones están muy actualizadas.

Este fichero al contener la puntuación y el número de votos de cada contenido tiene relación con el objetivo de analizar las producciones más populares y nos permitirá comprobar si se cumplen los KPIs definidos en ese objetivo (KPI 2.1 y KPI 2.2).

A continuación, podemos ver la tabla de metadatos de title.rating.tsv, en donde podemos observar detalladamente cada campo del fichero.

Fichero	Nombre columnas	Tipo datos	Descripción	Ejemplo
title.ratings.tsv	Tconst	String	id de la película o serie en IMDb	tt0332452
	averageRating	Float	puntuación media de IMDb	7.3
	numVotes	Int	número de votos	574440

Tabla 2. Metadatos title.rating.tsv

6.1.1.3 name.basics.tsv

Este fichero contiene 1301485 filas de información de profesionales del cine desde sus inicios hasta la actualidad. Incluye datos originales de IMDb. Este fichero se actualiza diariamente por lo que la información de este fichero está muy actualizada.

A continuación, podemos ver la tabla de metadatos de name.basics.tsv, en donde podemos observar detalladamente cada campo del fichero.

Fichero	Nombre columnas	Tipo datos	Descripción	Ejemplo
name.basics.tsv	Nconst	String	id de la persona	nm0000093
	primaryName	String	nombre de la persona	Brad Pitt
	birthYear	Int	año nacimiento de la persona	1963
	deathYear	Int	año muerte de la persona	\N
	primaryProfession	String	profesiones de la persona	producer,actor,executive
	knownForTitles	String	títulos en los que participa	tt0137523,tt0356910,tt1210166, tt0114746

Tabla 3. Metadatos name.basics.tsv

6.1.2 IMDb – webscraping

Cabe destacar que los ficheros que se descargaron de la sección "Non Commercial Datasets" de IMDb, no se usarán tal cual para realizar el análisis. Estos ficheros descargados se sobrescribirán para añadirles campos adicionales y enriquecerlos para que el análisis que realicemos sea más completo. Previamente se eliminaron todos aquellos contenidos de los ficheros mencionados anteriormente que no sea una película o serie.

Para enriquecer los ficheros base, se utilizarán técnicas de webscraping para obtener información adicional tanto de los contenidos, como de profesionales de la industria, estas técnicas de webscraping se explicarán más detalladamente en la sección de ETL. Dicha información se obtendrá de la página de IMDb tanto de información de la sección de contenidos audiovisuales como de profesionales de la industria (actores, directores, guionistas).

La información de la página de IMDb se va actualizando a diario, por lo que los datos son muy fiables.

6.1.2.1 title.basics.tsv

En este caso con técnicas de webscraping añadiremos información adicional de los contenidos audiovisuales. Los datos los extraeremos en formato tabular, posteriormente a los datos que venían previamente en el fichero base.

Este fichero contiene información sobre el tipo de contenido y país de producción, lo que está relacionado con el objetivo de analizar las producciones nacionales y nos permite ver si se cumplen los KPIs definidos de ese objetivo (KPI 1.1 y 1.2). También contiene información sobre las fechas de lanzamiento de los contenidos, lo que está relacionado con el objetivo de analizar los lanzamientos en fechas clave de España y EE. UU y nos permite comprobar si se cumplen los KPIs de ese objetivo (KPI 3.1, 3.2, 3.3 y 3.4). Además, el fichero contiene los diferentes premios Oscar que gano una película, lo que está relacionado con el objetivo de analizar el reconocimiento en premios Oscar de los contenidos y nos permite comprobar si se cumplen o no los KPIs definidos en este objetivo (KPI 5.1 y 5.2).

Los datos que se obtendrán por webscraping son las nominaciones, los premios, el presupuesto, la recaudación, la empresa productora, la fecha de lanzamiento, el país de

producción, los actores, el director, el guionista y si el contenido gano algún Oscar o no. A continuación, podemos ver los datos extraídos en formato tabla.

Fichero	Nombre columnas	Tipo datos	Descripción	Ejemplo
titles.basics.tsv	Awards	Int	cantidad de premios	5
	Nominations	Int	cantidad de nominaciones	23
	Budget	String	presupuesto	$175,000,000 (estimated)
	grossReveune	String	recaudación mundial	$497,409,852
	productionCompany	String	empresa productora	Warner Bros.
	releaseDate	Date	fecha de lanzamiento (dd/mm/aaaa)	14/05/2004
	Country	String	país del cual es la película o serie	United States
	Cast	String	actores principales	nm0000093, nm0051509, nm0089217
	oscarWon	Array String	premios Oscar ganados	[ ]

Tabla 4. Datos extraídos por webscraping sobre contenidos

6.1.2.2 name.basics.tsv

Mediante webscraping añadiremos información adicional de los profesionales de la industria. Los datos se pretenden extraer en formato tabular, añadiendo los siguientes campos al fichero original: el lugar de nacimiento (país, estado y ciudad), fecha de nacimiento, premios ganados, y si la persona ha ganado algún Oscar. A continuación, podemos ver un ejemplo de cómo se muestran parte de los datos en IMDb y su extracción en formato tabla.

Nacimiento 18 de diciembre de 1963 · Shawnee, Oklahoma, Estados Unidos

Figura 27. Datos de nacimiento de un actor a extraer - imdb.com

Fichero	Nombre columnas	Tipo datos	Descripción	Ejemplo
name.basics.tsv	Birth Date	Date	fecha nacimiento de la persona	18/12/1963
	City of Birth	String	ciudad nacimiento de la persona	Shawnee
	State of Birth	String	estado de nacimiento de la persona	Oklahoma
	Country of Birth	String	país nacimiento de la persona	Estados Unidos
	Awards Count	Int	cantidad de premios ganados	122
	Won Oscar	Boolean	indica si la persona gano algún Oscar	True

Tabla 4. Datos extraídos por webscraping sobre profesionales

6.1.3 JustWatch

Posteriormente, para conocer la disponibilidad de contenidos en plataformas de streaming, se obtuvieron los datos de JustWatch que es una web la cual funciona como una guía de streaming gratuita, esta permite ver que contenidos hay en numerosas plataformas de streaming, nos centraremos en obtener la información de contenidos que pertenezcan a alguna de estas tres plataformas: Netflix, Prime Video y Apple TV. Los contenidos de JustWatch también se actualizan diariamente, por lo que la fiabilidad es alta. Para obtener todos los datos se utilizó también técnicas de webscraping, dicha información se guardará en los ficheros netflix.csv, prime.csv, y appletv.csv, ficheros que son generados desde cero.

6.1.3.1 netflix.csv, prime.csv y appletv.csv

Cada uno de los tres ficheros contiene los títulos de películas y series disponible en la plataforma correspondiente extraído de JustWatch mediante técnicas de webscraping.

El fichero netflix.csv contiene 1924 filas, prime.csv contiene 1814 filas y appletv.csv contiene 245 filas. Los contenidos de JustWatch abarcan de 1974 hasta 2025, además lo actualizan diariamente, y los datos de estos ficheros se obtuvieron en mayo, por lo que están muy actualizados.

El fichero netflix.csv contiene información sobre los contenidos que están disponibles en esa plataforma, lo que está relacionado con el objetivo de analizar los contenidos de Netflix y nos permite ver si se cumplen o no los KPIs definidos en ese objetivo (KPI 4.1, 4.2 y 4.3).

A continuación, podemos ver la tabla de metadatos de los 3 ficheros en donde podemos observar detalladamente cada campo del fichero.

Ficheros	Nombre columnas	Tipo datos	Descripción	Ejemplo
netflix.csv	title_spanish	String	título en español de la película o serie	El juego del calamar
	Tconst	String	id de la película o serie en IMDb	tt10919420
prime.csv	Actors	String	actores de la película o serie	nm0497631, nm1059827, nm1601080
	Director	String	director de la película o serie	
appletv.csv	Writrer	String	escritor de la película o serie	nm1375363

Tabla 6. Metadatos netflix.csv, prime.csv y appletv.csv

Los tres ficheros tienen los mismos campos: title_spanish, tconst, actors, director y writter. Todos estos campos los hemos obtenido mediante webscraping de la página JustWatch.

6.2 Diseño de almacén de datos

6.2.1 Diseño conceptual

A continuación, lo siguiente que haremos es hacer el diseño conceptual de nuestro almacén de datos. Este diseño es importante para definir las funciones y características clave de nuestro sistema. Para realizarlo, podemos usar cualquier herramienta que nos permite crear diagramas, en mi caso utilizare draw.io, la cual nos ofrece una herramienta online que nos permite crear diagramas de forma gratuita, es muy intuitiva y fácil de utilizar. El diagrama lo podemos ver a continuación.

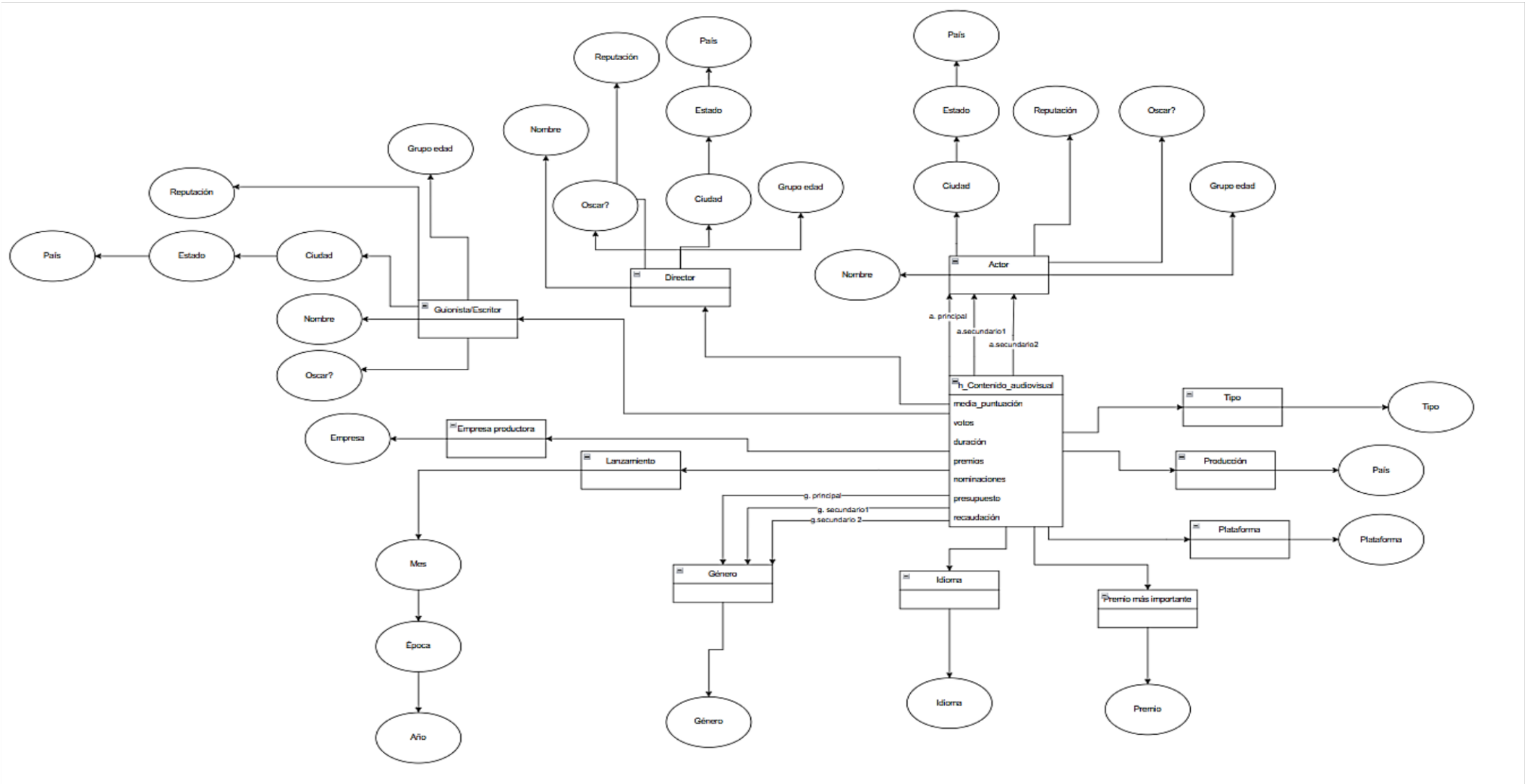


Figura 28. Diseño conceptual - Elaboración propia

En el diagrama conceptual, como podemos ver en la imagen, no contiene aún tipos, ni campos. Este diseño sirve para ver una primera impresión de como quedaría nuestro almacén de datos.

En primer lugar, tenemos la entidad central de nuestro esquema estrella que es "Contenido audiovisual". Lo llamo de esta manera, ya que engloba tanto películas como series. En esta entidad encontramos las medidas del proyecto, como el presupuesto o la recaudación de un contenido. También encontramos las diferentes claves ajenas que relaciona cada contenido con las diferentes dimensiones.

A continuación, encontramos la dimensión "Lanzamiento", que permite contextualizar temporalmente los diferentes contenidos, lo que nos permitirá realizar un análisis histórico con las diferentes fechas de lanzamiento y está relacionado con el objetivo de analizar los lanzamientos en fechas clave y nos permitirá comprobar si se cumplen los KPIs definidos en ese objetivo (KPI 3.1, 3.2, 3.3 y 3.4). En esta dimensión, se incluyen las jerarquías mes, época y año en el que se estrenó el contenido audiovisual. En época incluiremos las diferentes estaciones del año además de Navidad para saber si una película se estrena en esa época.

Por otro lado, encontramos la dimensión "Género", que nos permitirá realizar análisis históricos sobre los diferentes géneros, así también como poder observar su evolución en el tiempo. Además, esta dimensión servirá para comprobar si se cumple el KPI 2.2 del objetivo de analizar las producciones más populares y el KPI 4.3 del objetivo de analizar los contenidos de Netflix.

Como podemos ver en nuestro esquema si tenemos varios géneros nos interesa quedarnos con tres (uno principal y dos secundarios).

También encontramos otras dimensiones sobre los contenidos audiovisuales como "Empresa productora", que es la empresa que llevo a cabo.

Además, encontramos la dimensión "Idioma" que nos permite saber si un contenido audiovisual, únicamente está en un idioma o si tiene varios idiomas disponibles.

A continuación, la dimensión "Tipo" nos indica el tipo de la producción audiovisual que será película o serie. Esta dimensión prácticamente está relacionada con todos los objetivos definidos.

Por otro lado, la dimensión "Producción" que nos indica el país en el que se produjo el contenido audiovisual. Esta información está relacionada con varios objetivos, como el de analizar las producciones de cine nacional, que nos permite comprobar si se cumplen los KPIs definidos en este objetivo (KPI 1.1 y 1.2), o también el objetivo de analizar los lanzamientos en EE. UU y España, para poder ver si se cumplen los KPIs de este objetivo (KPI 3.1, 3.2, 3.3 y 3.4).

La dimensión "Plataforma", nos permite saber si algún contenido audiovisual, está en Netflix, AppleTV+, o Prime Video. La información de la dimensión está relacionada con el objetivo de realizar un análisis de los contenidos de Netflix y nos permite comprobar si se cumplen los KPIs definidos en dicho objetivo (KPI 4.1, 4.2 y 4.3).

La dimensión "Premio más importante" permite conocer el premio más importante que gano un contenido audiovisual. La información de la dimensión está relacionada con el objetivo de analizar el reconocimiento de los premios Oscar de los contenidos, y nos permite ver si se cumplen los KPIs establecidos en ese objetivo (KPI 5.1, 5.2).

Por último, tenemos las dimensiones "Actor", "Director" y "Escritor/Guionista", de los cuales podemos ver información como su nombre, grupo de edad, si gano un Oscar y lugar de nacimiento.

6.2.2 Diseño lógico

Ahora que ya hemos analizado nuestras fuentes de datos y hemos creado el diseño conceptual, procedemos a realizar el diseño lógico del almacén de datos. Este diseño se basa en la creación de un modelo que define la estructura de los datos y las relaciones entre ellos. Además, se utilizan herramientas como diagramas de entidad-relación, diagramas de flujo de datos y diagramas de clase para representar la estructura del sistema. Para realizarlo he utilizado el software MySQL Workbench. A continuación, podemos ver la imagen de mi diseño lógico.

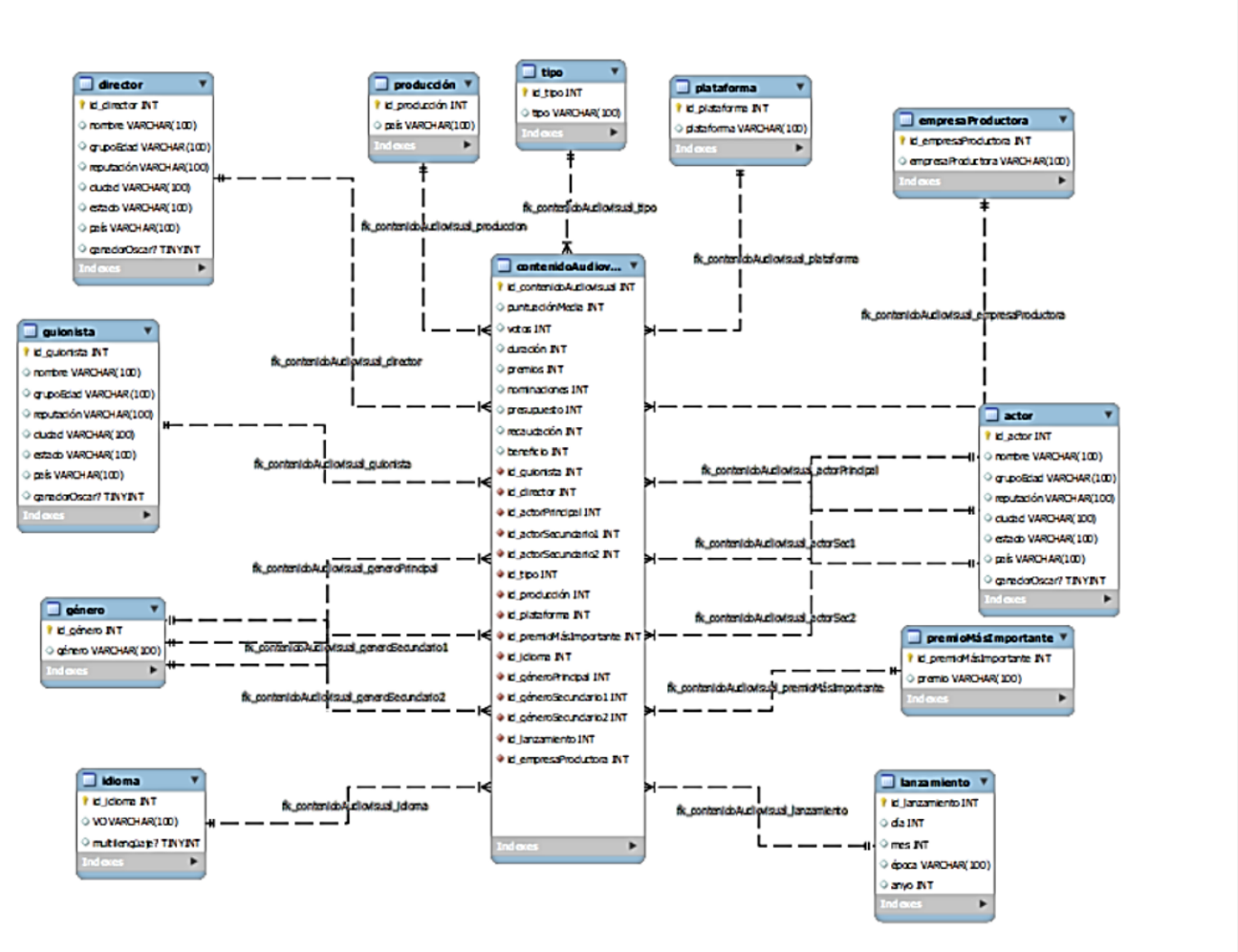


Figura 29. Diseño lógico - Elaboración propia

Una vez creado el diseño lógico, MySQL WorkBench nos da la opción de crear un script para poder crear nuestro almacén de datos, este se conoce como diseño físico. Una vez obtenido ese script debemos realizar unos ajustes para poder adaptarlo al lenguaje de Hive, ya que Hive, tiene algunas peculiaridades que explicamos anteriormente.

Para poder ver el script que nos permite desplegar las tablas en la base de datos del almacén de datos podemos verlo desde mi repositorio de GitHub en la carpeta "diseños" [52]. Para adaptarlo al lenguaje de Hive, se eliminaron las claves ajenas ya que no lo permite. También hay algunos tipos de datos no son iguales en el script generado y en Hive, por ejemplo:

- VARCHAR(n) se sustituyó por STRING, ya que en Hive no hay que indicar una longitud fija.

- TINYINT se sustituyó por BOOLEAN, ya que Hive no lo soporta.

Una ejecutado el script, ya podemos comprobar que las tablas se han creado correctamente.

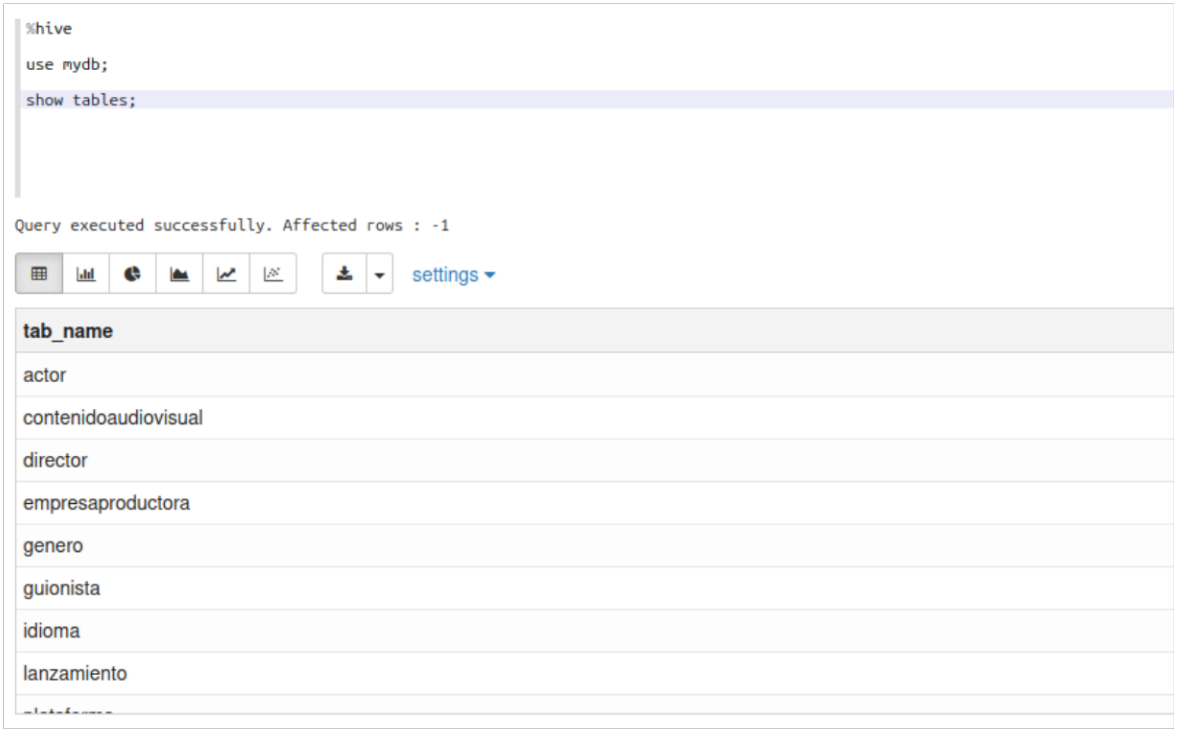


Figura 30. Tablas en la base de datos de Hive - Elaboración propia

6.3 Procesos ETL

Como se mencionó anteriormente, para realizar los ETL utilizaremos el lenguaje de programación Scala que es lenguaje por defecto que viene con Spark. El código completo de todos los ETLs podemos verlo en mi repositorio GitHub en la carpeta "etls" [53]. En el repositorio se encontrarán los scripts completos en Scala para Spark en Zeppelin.

El diseño de los ETL incluye la extracción de datos desde los ficheros, su transformación y limpieza, y finalmente su carga en las tablas de Hive.

A continuación, se explicarán como se obtuvieron los ficheros resultantes que serán contendrán la información necesaria para realizar los ETL.

6.3.1 Extracción de datos por webscraping

Se utilizaron técnicas de webscraping tanto para sobrescribir y enriquecer los ficheros públicos descargados de IMDb en la sección "Non Commercial Datasets", como para crear

ficheros desde cero que contiene información sobre diferentes contenidos audiovisuales que pertenecen a alguna plataforma de streaming de la página JustWatch. Los scripts están programados en Python y los podemos encontrar en mi repositorio de GitHub en la carpeta "webscraping" [53].

6.3.1.1 IMDb

IMDB nos ofrece ficheros oficiales con millones de datos de contenidos audiovisuales bastante completos, no obstante, no aparecían en estos datasets información como el presupuesto, la recaudación, o detalles de premios.

Para obtener esta información que no aparece en los ficheros originales, se crearon dos scripts, uno para obtener información adicional para los contenidos audiovisuales, como el presupuesto, la recaudación, y el año de lanzamiento y otro script para obtener información adicional sobre los actores, directores y guionistas, como país, estado, ciudad y fecha de su nacimiento. Dicha información se obtendrá de la página oficial de IMDb según el contenido o la persona que corresponda, ya que cada uno se identifica con un id.

Los scripts leen de uno en uno los identificadores del contenido o de la persona del dataset, descarga la información necesaria de la página de IMDb mediante webscraping aplicando funciones de la librería BeautifulSoup.

Se utilizaron también diferentes threads o peticiones, para agilizar el proceso ya que si procesaba de manera individual cada petición no iba a dar tiempo a procesar toda la información ya que perfectamente podía tardar años procesar toda la información. Al principio tardaba mucho y a veces había errores a la hora de obtener los datos y me bloquearon la ip temporalmente por hacer muchas peticiones en poco tiempo, ya que este tipo de páginas tienen programas que detectan comportamiento sospechoso como ataques DDoS.

Para evitar que me boquearan la ip, se añadieron pequeñas pausas aleatorias entre la obtención de datos de cada petición y se usaron diferentes "user-agents" los cuales son programas informáticos que simulan diferentes usuarios en la web con diferente ip. De esta manera se consiguió obtener la información de manera más rápida sin ser bloqueado. Cada vez que se obtuvo la información de un bloque de 1000 contenidos o profesionales se iba sobrescribiendo sobre la información que ya se había scrapeado anteriormente, de modo que

si hay un fallo puedo continuar desde donde lo dejé sin perder todo lo que se scrapeo. En total extraer esta información adicional mediante webscraping llevo varios meses.

```
Procesado nm0000005: Ingmar Bergman, 14/07/1918, Uppsala, Uppsala län, Suecia, Premios: 81, Oscar: False
Procesado nm0000002: Lauren Bacall, 16/09/1924, Bronx, Nueva York, Estados Unidos, Premios: 23, Oscar: F
alse
Procesado nm0000001: Fred Astaire, 10/05/1899, Omaha, Nebraska, Estados Unidos, Premios: 33, Oscar: Fals
e
Procesado nm0000008: Marlon Brando, 03/04/1924, Omaha, Nebraska, Estados Unidos, Premios: 36, Oscar: Tru
e
Procesado nm0000015: James Dean, 08/02/1931, Marion, Indiana, Estados Unidos, Premios: 8, Oscar: False
Procesado nm0000006: Ingrid Bergman, 29/08/1915, None, Estocolmo, Suecia, Premios: 50, Oscar: True
Procesado nm0000014: Olivia de Havilland, 01/07/1916, None, Tokio, Japón, Premios: 28, Oscar: True
Procesado nm0000011: Gary Cooper, 07/05/1901, Helena, Montana, EE.UU., Premios: 43, Oscar: True
Procesado nm0000012: Bette Davis, 05/04/1908, Lowell, Massachusetts, Estados Unidos, Premios: 56, Oscar:
 True
Procesado nm0000010: James Cagney, 17/07/1899, None, Nueva York, Estados Unidos, Premios: 31, Oscar: Tru
e
```

Figura 31. Scrapeo información adicional profesionales - Elaboración propia

6.3.1.2 JustWatch

La página JustWatch no ofrece datasets públicos con su información de películas y series en diferentes plataformas, por los ficheros que usaremos de los contenidos que pertenecen a las plataformas de streaming Netflix, Prime Video o Apple TV + lo obtendremos mediante webscraping, guardando los contenidos en diferentes csv que generaremos nosotros según al contenido al que pertenezca. En dichos ficheros se separa cada campo de información con comas.

Al igual que con IMDb se utilizaron tiempos de espera, varios threads, pequeñas pausas aleatorias, para agilizar el proceso y evitar que nos bloqueen la ip.

Para obtener esta información, además de usar funciones de la librería de BeautifulSoup, también se utilizará Selenium WebDriver, una librería la cual nos permite simular acciones de usuario, esto porqué al cargar la página de JustWatch no aparecen directamente la información de todos los contenidos, para ello hay que scrollear la pantalla hacia abajo. Entonces para obtener los datos mediante Selenium WebDriver, primero se abrirá un navegador en Google Chrome accediendo a la página de JustWatch de la plataforma de streaming correspondiente. Posteriormente se simula un scroll de usuario con Selenium WebDriver y este va bajando poco a poco para cargar los contenidos y posteriormente obtener su información mediante funciones de la librería BeautifulSoup. Una vez finalizado el scrapeo se cerrará el navegador y se guardará los datos en el csv correspondiente.

Los contenidos también se actualizan diariamente, por lo que la fiabilidad es alta y la información fue obtenida a finales de mayo.

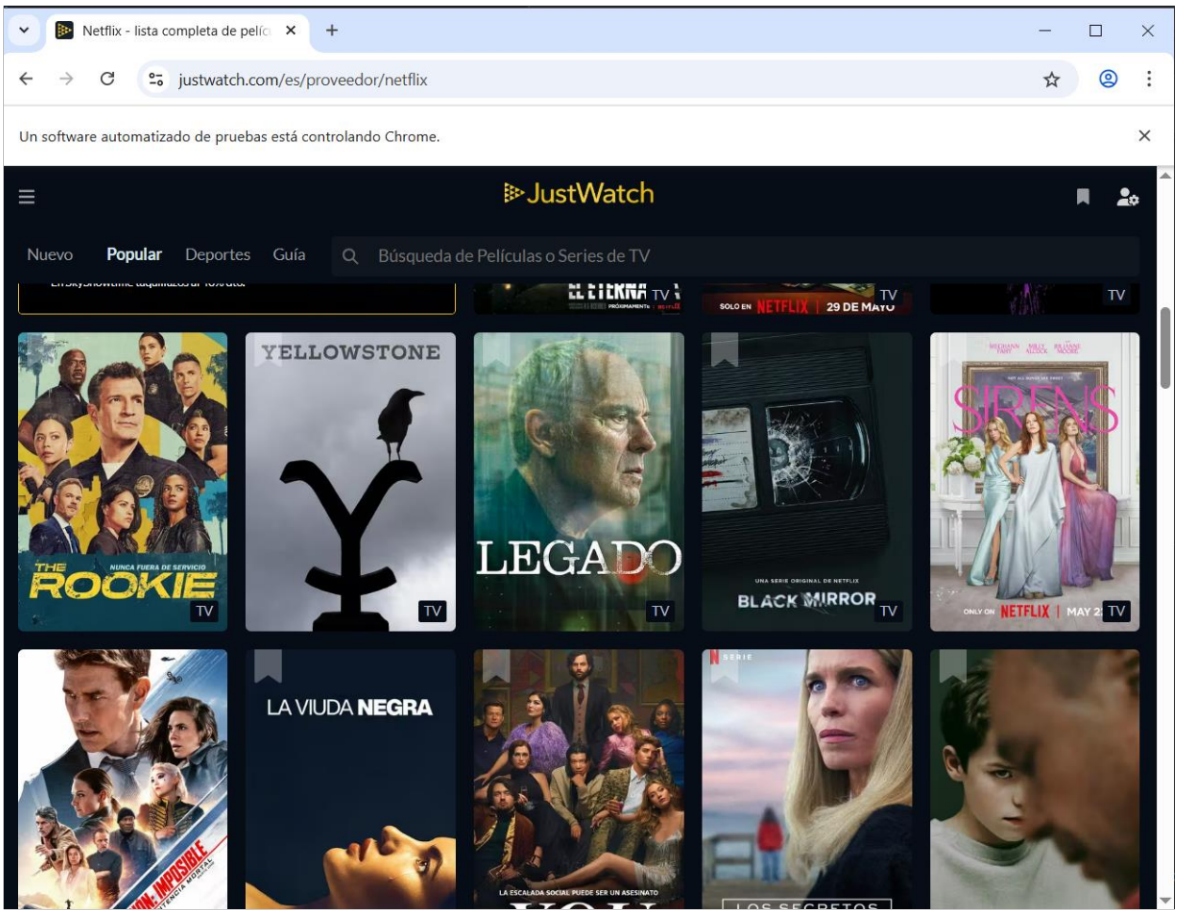


Figura 32. Scrolleo simulado por Selenium WebDriver de contenidos de Netflix - justwatch.com

```
Extrayendo títulos:
Película 1: El Eternauta
Película 2: Department Q
Película 3: Love, Death & Robots
Película 4: El Novato
Película 5: Yellowstone
Película 6: Legado
Película 7: Black Mirror
Película 8: Sirenas
Película 9: Misión: Imposible - Sentencia mortal parte uno
Película 10: La viuda negra
Película 11: You
Película 12: Los secretos que ocultamos
Película 13: Adolescencia
Película 14: El contable
Película 15: The Four Seasons
```

Figura 33. Scrapeo información adicional contenidos en plataformas – Elaboración propia

6.3.2 Transformación y carga de datos

6.3.2.1 Dimensión tipo

En la dimensión tipo, solo nos interesa tener dos posibles tipos de contenido audiovisual en nuestro almacén de datos, las películas y las series. En este caso no es necesario usar ningún fichero base para cargar los datos en la tabla del almacén de datos.

Esta dimensión prácticamente está relacionada con todos los objetivos definidos, ya por ejemplo en el objetivo de analizar las producciones nacionales, se estableció un KPI para comprobar si en los últimos años hubo un aumento de películas españolas KPI 1.1 y otro para series españolas KPI 1.2.

En el dataset title.basics.tsv aparecen otros posibles tipos como shows de tv, cortometrajes, videojuegos, los cuales no corresponden a la definición de contenido audiovisual. Por tanto, en este dataset se decidió eliminar aquellos contenidos que no sean películas o series, es decir que no sean "movie", "tvMovie" o "tvSerie".

En nuestra tabla Hive tendremos 2 tipos posibles, Película para los contenidos que como tipo en el dataset tiene "movie" o "tvMovie" o Serie para los contenidos que como tipo en el dataset tiene "tvSerie".

Una vez realizado el ETL, con Zeppelin podemos visualizar como quedaría nuestra tabla tipo en el almacén de datos.

settings

id_tipo	tipo
1	Pelicula
2	Serie

Figura 34. Tabla tipo - Elaboración propia

6.3.2.2 Dimensión plataforma

Para esta dimensión nos interesa guardar las posibles plataformas a las que puede pertenecer cada contenido audiovisual. Los datasets de los contenidos que pertenecen a alguna plataforma, como se dijo anteriormente se obtuvieron por webscraping en donde había que simular acciones de usuario al scrapear la información con la librería Selenium WebDriver,

ya que no se cargaba toda la información al acceder a la página web de JustWatch que es de donde obtuvimos la información.

Cada contenido puede pertenecer a ninguna, una o varias plataformas y las posibles plataformas son Netflix, AppleTV+ y Prime Video.

La información de la dimensión está relacionada con el objetivo de realizar un análisis de los contenidos de Netflix y nos permite comprobar si se cumplen los KPIs definidos en dicho objetivo (KPI 4.1, 4.2 y 4.3).

Los posibles valores que tendremos en la tabla Hive plataforma son: Sin plataforma, para aquellos contenidos que no pertenecen a ninguna de las tres plataformas de streaming, Multiplataforma, si el contenido pertenece a más de una plataforma de streaming, Netflix, AppleTV+, o Prime Video.

Hubo problemas a la hora de obtener la información de JustWatch, ya que algunos títulos estaban duplicados y había algunas entradas vacías porque no se encontró esa película en IMDb, no se pudo obtener el tconst (código de esa película en IMDb), por lo que estas películas se eliminaron de los datasets de las plataformas de streaming, ya que sin ese código posteriormente no se podrá hacer los joins correspondientes. Se aplicó un proceso para eliminar duplicados y validar que los contenidos que scrapeamos de JustWatch se encuentra su código en IMDb para poder hacer los joins.

La información de las películas que pertenecen algunas plataformas la encontramos en los ficheros netflix.csv, primevideo.csv y appletv.csv aunque en esta dimensión no es necesario utilizarlo ya que únicamente estamos cargando el nombre de la plataforma. Estos ficheros nos serán útiles posteriormente en la tabla de hechos para hacer los correspondientes joins.

Para la carga en Hive se insertaron cinco valores fijos, asignándoles ids de 1 a 5.

settings

id_plataforma	plataforma
1	Sin plataforma
2	Multiplataforma
3	Netflix
4	PrimeVideo
5	AppleTVPlus

Figura 35. Tabla plataforma - Elaboración propia

6.3.2.3 Dimensión producción

En la tabla de producción, cargaremos los diferentes países en el que se realizaron cada contenido audiovisual. Para ello los obtendremos del fichero "title.basics.tsv". Aquí tenemos que cargar la columna country, y encontramos el problema de que puede haber registros mal extraídos, por lo que realizaremos una limpieza de nombres de país con caracteres especiales como acentos. Además, que los mismos países podían estar en diferentes idiomas (España, Spain), minúsculas y mayúsculas mezcladas. Para que todos los países tuvieran el mismo formato se utilizó en primer lugar una función que pasaba el texto a minúsculas y eliminaba caracteres especiales como acentos. Posteriormente, se hacen llamadas al API de Google Translate para traducir cualquier idioma en el que este el país al inglés. Tras la conversión, se pasa la primera letra a mayúsculas. De esta manera nos queda normalizado todos los países en inglés.

Por otro lado, para los casos en lo que los países de producción de un contenido audiovisual no se conozcan (sea null), creamos una producción especial que se llame "Desconocido" y le asignamos el id = 1, lo asignamos el primero para que se vea más claro. Luego para el resto de los países se le asigno un id incremental, además de previamente haber hecho la limpieza de datos comentada anteriormente.

La información de esta dimensión está relacionada con varios objetivos, como el de analizar las producciones de cine nacional, que nos permite comprobar si se cumplen los KPIs definidos en este objetivo (KPI 1.1 y 1.2), o también el objetivo de analizar los lanzamientos en EE. UU y España, para poder ver si se cumplen los KPIs de este objetivo (KPI 3.1, 3.2, 3.3 y 3.4).

settings

id_produccion	pais
1	Desconocido
2	Afghanistan
3	Albania
4	Algeria
5	American Samoa
6	Andorra
7	Angola
8	Anguilla

Figura 36. Tabla producción - Elaboración propia

6.3.2.4 Dimensión empresa productora

Este ETL se realiza de manera muy similar al anterior. Obtendremos la información del fichero title.basics.tsv y extraeremos el campo 'productionCompany'. También hubo que realizar una limpieza de datos, ya que las empresas podían aparecer abreviadas, mal escritas o repetidas con nombres parecidos ("Warner Bros", "Warnes Bros Pictures), mayúsculas/minúsculas, caracteres especiales.

Para solucionar en primer lugar se creó una función que elimina caracteres especiales y pasa el texto a minúsculas. Posteriormente, con otra función, se agruparán aquellas entradas cuya versión normalizada coincida parcialmente en su contenido, de modo que únicamente se conserve la variante de menor longitud de cada grupo. Por ejemplo, si tenemos "Warner Bros" "Warner-Bros" y "Warnes Bros Pictures", se guardará la de menor longitud que es "Warner Bros".

También hubo para los campos vacíos o null se asignará la empresa productora "Desconocida", que le asignaremos el id 1. Para el resto de las empresas se le asigno un id incremental.

A continuación, podemos ver la tabla en donde tenemos las diferentes empresas productoras diferentes, contando aquellas que son desconocidas.

settings

id_empresaproductora	empresaproductora
200993	Ten Plus
200994	Ten Point Oh Films
200995	Ten Seconds
200996	Ten Speed Media
200997	Ten Sundays Productions
200998	Ten TV
200999	Ten Ten Films

Figura 37. Tabla empresa productora - Elaboración propia

6.3.2.5 Dimensión género

Para obtener los diferentes géneros que puede tener un contenido audiovisual, los obtendremos del fichero title.basics.tsv. De manera similar a ETLs anteriores. Extraemos en este caso la columna 'genres'. Para los contenidos en los que no se conoce el género (es null), se creó un caso llamado "Desconocido" el cual se le asignara id = 1, para el resto de los géneros se asignara un id incremental.

Esta dimensión servirá para comprobar si se cumple el KPI 2.2 del objetivo de analizar las producciones más populares y el KPI 4.3 del objetivo de analizar los contenidos de Netflix.

Aquí el problema es que un contenido audiovisual puede tener más de un género, los géneros están en formato string separado por comas (si hay más de un género), y con un espacio en blanco entre ellos, por lo que se usó explode() para poder obtener los diferentes géneros de forma individual como podemos ver en el siguiente fragmento de código.

```
// Dividir el campo 'genres' por comas
val explodedDF = df.selectExpr("explode(split(genres, ',')) as genero")

// Eliminar espacios en blanco, valores nulos y duplicados
val cleanedDF = explodedDF.withColumn("genero", trim(col("genero")))
```

```
.filter(col("genero").isNotNull && col("genero") =!= "\\N")

.dropDuplicates()
```

settings

id_genero	genero
1	Sin genero
2	Action
3	Adult
4	Adventure
5	Animation
6	Biography
7	Comedy
8	Crime

Figura 38. Tabla género - Elaboración propia

6.3.2.6 Dimensión idioma

En esta dimensión queremos almacenar los diferentes idiomas que puede tener un contenido audiovisual, al igual que ETLs anteriores este campo se obtendrá del fichero title.basics.tsv. Se incluyo una categoría "Multilenguaje" por si hay más de un idioma y "Desconocido" por si no hay ninguno (el campo "languajes" es null). Aquí el problema es que en el dataset al scrappear el campo idioma, en la página de IMDb hay contenidos audiovisuales que solo aparece como idioma la versión original (aunque realmente tiene más idiomas), entonces he decidido que en esos casos lo trataremos como si tuvieran un único idioma.

Además, hubo que hacer una limpieza ya que podían aparecer en diferentes idiomas (Spanish, español), por lo que se añadieron funciones de traducción como se hizo en otros ETLs.

En el caso "Multilenguaje" se asignará el id 1, para los "Desconocido" se asigna el id 2 y para el resto de los idiomas se asigna un id incremental.

Esta dimensión será necesaria para más adelante responder a la siguiente pregunta analítica: ¿Qué idiomas originales son los más frecuentes en los contenidos?

settings

id_idioma	idioma
1	Multilenguaje
2	Desconocido
3	Abkhazian
4	Aboriginal
5	Acholi
6	Afrikaans
7	Akan
8	Albanian

Figura 39. Tabla idioma - Elaboración propia

6.3.2.7 Dimensión premio más importante

En esta dimensión vamos a cargar los posibles premios Oscar que puede ganar un contenido audiovisual. Posteriormente en tabla de hechos, decidiremos el orden de importancia de estos premios Oscar.

La información de la dimensión está relacionada con el objetivo de analizar el reconocimiento de los premios Oscar de los contenidos, y nos permite ver si se cumplen los KPIs establecidos en ese objetivo (KPI 5.1, 5.2).

En caso de que una película tiene varios premios se considera solo el que tenga mayor importancia. Los premios los obtenemos de la columna "oscarsWon" del fichero title.basics.tsv.

Para está dimensión también se creó un caso "Sin premio" para aquellos contenidos que no ganaron ningún premio Oscar. Además, para esta dimensión hubo que realizar una limpieza mediante una función regex, ya que el nombre de los premios aparecía entre corchetes de forma anidada, separado por comas, con comillas simples, además de que había premios que eran el mismo premio, pero algunos estaban en mayúsculas, abreviados, por lo que también hubo que limpiar esa información con algunas funciones. A continuación, podemos ver en el siguiente fragmento de código como realizamos esa limpieza de datos.

```
// Limpieza: quitar corchetes y comillas simples

val dfPremiosClean = dfPremiosRaw.withColumn("premio",
```

```
  regexp_replace(regexp_replace(col("premio"),      "\\[|\\]",
""), "'", ""))


// Dividir por comas en caso de tener más de un premio

val  dfPremiosExploded  =  dfPremiosNorm.withColumn("premio",
explode(split(col("premio"), ",")))
```

settings

id_premiomasimportante	premio
1	Sin premio
2	Best Achievement in Art Direction
3	Best Achievement in Cinematography
4	Best Achievement in Costume Design
5	Best Achievement in Directing
6	Best Achievement in Film Editing
7	Best Achievement in Makeup
8	Best Achievement in Makeup and Hairstyling

Figura 40. Tabla premio más importante - Elaboración propia

6.3.2.8 Dimensión lanzamiento

En esta dimensión se generarán todas las fechas (mes – época – año) que se estrenan los contenidos audiovisuales en España.

Esta dimensión está relacionada con el objetivo de analizar los lanzamientos en fechas clave y nos permitirá comprobar si se cumplen los KPIs definidos en ese objetivo (KPI 3.1, 3.2, 3.3 y 3.4).

Las fechas se generarán mediante una función la cual genera las fechas desde 1984 hasta 2031 (hay información de contenidos que no se han estrenado), el cual es el intervalo de tiempo que se encontrarán los contenidos.

A partir del mes obtenemos la época en la que se estrena el contenido. Cabe destacar que tendremos 5 posibles épocas: Invierno (desde enero hasta marzo), Primavera (desde abril hasta junio), Verano (desde julio hasta septiembre), Otoño (desde octubre hasta noviembre) y Navidad (para el mes de diciembre). Aquí el problema es que muchos contenidos

audiovisuales no se conoce la fecha de lanzamiento, por lo que se creó un id de lanzamiento "Desconocido" para estos casos en donde el mes y el año será null y le asignaremos el id 1.

settings

id_lanzamiento	mes	epoca	anyo
1	null	Desconocido	null
2	1	Invierno	1903
3	1	Invierno	1910
4	1	Invierno	1911
5	1	Invierno	1912
6	1	Invierno	1913
7	1	Invierno	1914

Figura 41. Tabla lanzamiento - Elaboración propia

6.3.2.9 Dimensión actor

En esta dimensión cargaremos la información de los diferentes actores que participan en las producciones audiovisuales, dicha información la obtenemos del fichero name.basics.tsv. Entre dicha información, encontramos su nombre, su lugar de nacimiento (país, estado, ciudad), premios ganados, nominaciones, etc.

Esta dimensión nos servirá para responder a algunas preguntas analíticas como: ¿De qué países son los actores más populares?, y ¿A qué grupo de edad pertenecen los actores más conocidos?

A la hora de obtener el país, el estado y la ciudad se han tenido que realizar limpieza de datos, ya que aparecía el nombre de los países en diferentes idiomas, en algunos casos llevaba acentos, por lo que realizaron diferentes funciones de limpieza como en anteriores ETL se dejaron todos en inglés.

Es importante destacar que en el fichero name.basics.tsv, tenemos información de todas las personas involucradas en la industria cinematográfica, es decir que además de actores, habrá directores, guionistas, productores y otras profesiones relacionadas con el mundo audiovisual. De estas profesiones nos interesan los actores, es decir que en el campo "primaryProfession" aparezca actor o actress, por lo que habrá que realizar un filtrado por ese campo.

Originalmente el dato de la reputación de un actor no está en el fichero name.basics.tsv. Para determinar la reputación de un actor en nuestro almacén de datos, he realizado un histograma, en el cual podemos ver de manera gráfica la cantidad de premios que han ganado

los actores en forma de porcentaje. En total tenemos aproximadamente 900000 actores y cómo podemos ver la gran mayoría son desconocidos, es decir no han ganado ningún premio.

Además, posteriormente en la tabla de hechos, almacenaremos el id del actor principal, y los id de dos actores secundarios. Hay algunos contenidos en los que no se conoce alguno de los actores secundarios, u otros casos directamente no se conoce ningún actor, por lo que se creó un caso especial en almacén de datos llamado "Sin actor" para estos casos, con el resto de los campos nulos.

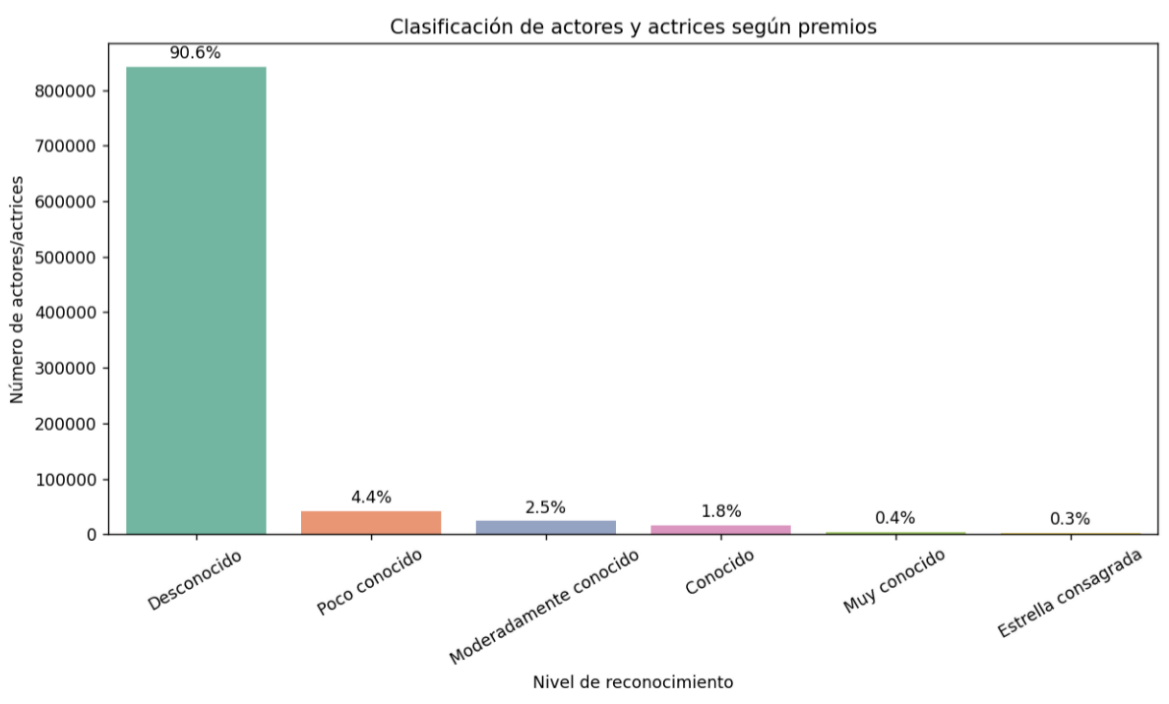


Figura 42. Histograma premios actores - Elaboración propia

Se puede ver la categorización de la reputación de los actores en la siguiente tabla:

Reputación	Premios ganados
Desconocido	0 premios
Poco conocido	1 premio
Algo conocido	2 – 3 premios
Conocido	4 – 10 premios
Muy conocido	10 – 20 premios
Estrella consagrada	+ 20 premios

Tabla 7. Categorización de actores por premios ganados

Por otro lado, para determinar el grupo de edad, el cual no viene dado directamente en el dataset, también vamos a categorizar. Cabe destacar que de la mayoría desconocemos el grupo de edad, porque esa información no está en IMDb. En primer lugar, se calculará la edad tomando su fecha de nacimiento y luego se le asignará un grupo de edad.

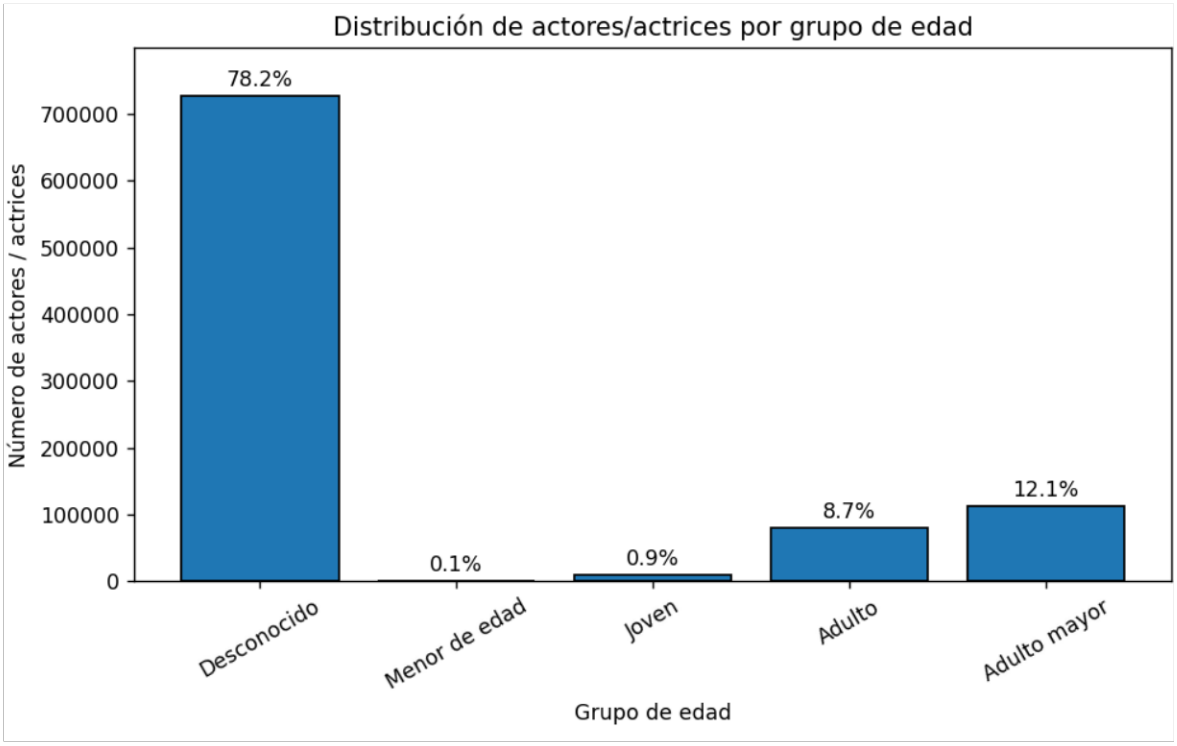


Figura 43. Histograma grupo edad actores - Elaboración propia

Se puede ver la categorización de los grupos de edad de los actores en la siguiente tabla:

Grupo edad	Edad
Desconocido	null
Menor de edad	0 – 17 años
Joven	19 – 30 años
Adulto	31 – 60 años
Adulto mayor	+ 60 años

Tabla 8. Categorización de actores por edad

id_actor	nconst	nombre	grupo_edad	reputacion	ciudad	estado	pais	ganadoroscar
1	n/a	Sin actor	null	null	null	null	null	null
2	nm0000001	Fred Astaire	Adulto mayor	Estrella consagrada	Omaha	Nebraska	United States	false
3	nm0000002	Lauren Bacall	Adulto mayor	Estrella consagrada	Bronx	New York	United States	false
4	nm0000003	Brigitte Bardot	Adulto mayor	Algo conocido	París	null	France	false
5	nm0000004	John Belushi	Adulto mayor	Algo conocido	Chicago	Illinois	United States	false
6	nm0000005	Ingmar Bergman	Adulto mayor	Estrella consagrada	Uppsala	Uppsala län	Sweden	false
7	nm0000006	Ingrid Bergman	Adulto mayor	Estrella consagrada	Estocolmo	null	Sweden	true
8	nm0000007	Humphrey Bogart	Adulto mayor	Estrella consagrada	New York City	New York	United States	true

Figura 44. Tabla actor - Elaboración propia

6.3.2.10 Dimensión director

En esta dimensión cargaremos la información de los diferentes directores que participan en las producciones audiovisuales, dicha información la obtenemos del fichero name.basics.tsv. Entre dicha información, encontramos su nombre, su lugar de nacimiento (país, estado, ciudad), premios ganados, nominaciones, etc.

A la hora de obtener el país, el estado y la ciudad se han tenido que limpiar datos, ya que aparecía el nombre de los países en diferentes idiomas, en algunos casos llevaba acentos, por lo que se utilizaron diferentes funciones de limpieza como en otros ETL y se dejaron todos en inglés.

Esta dimensión será necesaria para responder a la siguiente pregunta analítica más adelante: ¿Un director ganador de algún Oscar influye en la puntuación de la película?

Como se dijo anteriormente, en el fichero name.basics.tsv, además de directores, habrá otras profesiones relacionadas con el mundo audiovisual. De estas profesiones nos interesan los actores, es decir que en el campo "primaryProfession" aparezca director, por lo que habrá que realizar un filtrado por ese campo.

La categorización del campo grupo de edad y estado se hizo de la misma forma que en el ETL anterior.

Además, posteriormente en la tabla de hechos, almacenaremos el id del director principal. Hay algunos contenidos en los que no se conoce el director, por lo que se creó un caso especial en almacén de datos llamado "Sin director" para estos casos, con el resto de los campos nulos.

settings

id_director	nconst	nombre	grupo_edad	reputacion	ciudad	estado	pais	ganadoroscar
1	n/a	Sin director	null	null	null	null	null	null
2	nm0000005	Ingmar Bergman	Adulto mayor	Estrella consagrada	Uppsala	Uppsala län	Sweden	false
3	nm0000008	Marlon Brando	Adulto mayor	Estrella consagrada	Omaha	Nebraska	United States	true
4	nm0000009	Richard Burton	Adulto mayor	Muy conocido	Pontrhydyfen	Wales	United Kingdom	false
5	nm0000010	James Cagney	Adulto mayor	Estrella consagrada	New York City	New York	United States	true
6	nm0000018	Kirk Douglas	Adulto mayor	Estrella consagrada	Amsterdam	New York	United States	false
7	nm0000019	Federico Fellini	Adulto mayor	Estrella consagrada	Rimini	Emilia-Romagna	Italy	false
8	nm0000024	John Gielgud	Adulto mayor	Muy conocido	Londres	England	United Kingdom	true

Took 2 sec. Last updated by anonymous at May 02 2025, 10:04:44 AM. (outdated)

Figura 45. Tabla director - Elaboración propia

6.3.2.11 Dimensión guionista

En esta dimensión cargaremos la información de los diferentes guionistas involucrados en las producciones audiovisuales, dicha información la obtenemos del fichero name.basics.tsv. Entre dicha información, encontramos su nombre, su lugar de nacimiento (país, estado, ciudad), premios ganados, nominaciones, etc.

A la hora de obtener el país, el estado y la ciudad se han tenido que realizar limpieza de datos, ya que aparecía el nombre de los países en diferentes idiomas, en algunos casos llevaba acentos, por lo que se utilizaron diferentes funciones de limpieza como en otros ETL y se dejaron todos en inglés.

Como dijimos en las dimensiones anteriores, el fichero name.basics.tsv, contiene información de todas las personas involucradas en la industria cinematográfica. De estas todas las profesiones que hay nos interesan los guionistas, es decir que en el campo "primaryProfession" aparezca writer, por lo que habrá que realizar un filtrado por ese campo.

La categorización del campo grupo de edad y estado se hizo igual que en el ETL anterior.

Además, posteriormente en la tabla de hechos, almacenaremos el id del guionista principal. Hay algunos contenidos en los que no se conoce el guionista, por lo que se creó un caso especial en almacén de datos llamado "Sin guionista" para estos casos, con el resto de los campos nulos.

settings

id_guionista	nconst	nombre	grupo_edad	reputacion	ciudad	estado	pais	ganadoroscar
1	n/a	Sin guionista	null	null	null	null	null	null
2	nm0000004	John Belushi	Adulto mayor	Algo conocido	Chicago	Illinois	United States	false
3	nm0000005	Ingmar Bergman	Adulto mayor	Estrella consagrada	Uppsala	Uppsala län	Sweden	false
4	nm0000008	Marlon Brando	Adulto mayor	Estrella consagrada	Omaha	Nebraska	United States	true
5	nm0000019	Federico Fellini	Adulto mayor	Estrella consagrada	Rimini	Emilia-Romagna	Italy	false
6	nm0000024	John Gielgud	Adulto mayor	Muy conocido	Londres	England	United Kingdom	true
7	nm0000027	Alec Guinness	Adulto mayor	Estrella consagrada	Londres	England	United Kingdom	true
8	nm0000031	Katharine Hepburn	Adulto mayor	Estrella consagrada	Hartford	Connecticut	United States	true

Figura 46. Tabla guionista - Elaboración propia

6.3.2.12 Hecho contenido audiovisual

La tabla de hechos ha sido la parte más compleja de los ETLs, esto es debido a que hay una gran cantidad de claves ajenas, transformaciones de medidas, y tratamientos para casos especiales. El fichero que usaremos como base es title.basics.tsv.

En la tabla de hechos llamada contenidoAudiovisual, encontramos:

- Las claves ajenas relacionadas con las diferentes dimensiones.
- Las medidas: votos, duración, puntuacionMedia, presupuesto, recaudación y beneficio.
- El título y código de IMDb de la película (para poder hacer los respectivos joins)

Para obtener las medidas votos y puntuacionMedia, fue necesario hacer un join con el fichero title.ratings.tsv, ya que esos campos no aparecen directamente en el fichero que usamos como partida.

El beneficio es una medida calculada, la cual no venía directamente en el fichero base, la calculamos realizando la diferencia entre la recaudación menos el presupuesto.

En el fichero title.basics.tsv en la columna "titleType" encontramos tres posibilidades: movie, tvSeries, tvMovie, por lo que tendremos que realizar una normalización para transformar los tipos.

Cabe destacar que los campos presupuesto y recaudación, según el contenido audiovisual, podían aparecer en distintas monedas y formatos, por lo que se creó una función que realiza esos cambios de monedas y lo devuelve todo en euros en formato int.

En cuanto a las películas premiadas con premios Oscar, el premio más prestigioso de la industria cinematográfica celebrado en EE. UU, para relacionar un único premio (premio más importante) con un único contenido, se estableció un orden de importancia que podemos ver a continuación, de esta manera si un contenido tiene más de un premio, la clave ajena que apunta a esa dimensión será la del premio más importante. Todos estos premios son premios Oscar.

1. Best Picture → Mejor película
2. Best Director → Mejor dirección
3. Best Actor in a Leading Role → Mejor actor en un papel principal
4. Best Actress in a Leading Role → Mejor actriz en un papel principal
5. Best Original Screenplay → Mejor guion original
6. Best Adapted Screenplay → Mejor guion adaptado
7. Best Actor in a Supporting Role → Mejor actor de reparto
8. Best Actress in a Supporting Role → Mejor actriz de reparto
9. Best Cinematography → Mejor dirección fotográfica
10. Best Film Editing → Mejor montaje
11. Best Art Direction → Mejor dirección de arte
12. Best Costume Design → Mejor diseño de vestuario
13. Best Makeup and Hairstyling → Mejor maquillaje
14. Best Visual Effects → Mejor efectos visuales
15. Best Sound → Mejor sonido
16. Best Music (Original Score) → Mejor banda sonora original
17. Best Original Song → Mejor canción original
18. Best Foreign Language Film → Mejor película extranjera (fuera de EE. UU)
19. Best Animated Feature → Mejor película de animación
20. Best Documentary → Mejor Documental
21. Best Dance Direction → Mejor dirección de danza
22. Best Assistant Director → Mejor director asistente
23. Best Writing → Mejor guion

Para las claves ajenas de género (género principal, secundario1, secundario2) y actor (actor principal, secundario1 y 2), en caso de que un contenido no tuviera ningún género y ningún actor, todas las claves ajenas apuntarán al caso "Desconocido". Si solo se conoce uno, serían los secundarios los que apunten a "Desconocido" y así sucesivamente.

A continuación, podemos ver nuestra tabla de hechos tras realizar los joins necesarios.

id_contenidoaudiovisual	tconst	título	puntuacionmedia	votos	duracion	premios	nominaciones	presupuesto	recaudacion	beneficio
52828	tt0068646	The Godfather	9.2	2029674	175	31	31	5520000	230314822	224794822
52829	tt0068647	Gonshchiki	6.7	127	80	0	0	null	null	null
52830	tt0068648	Goodnight, My Love	6.7	180	73	0	0	null	null	null
52831	tt0068649	The Gore Gore Girls	5.2	2825	81	0	1	58420	null	null
52832	tt0068650	Hanzo the Razor: Sword of Justice	6.6	1909	94	0	0	null	null	null
52833	tt0068652	Gozenchû no jikanwari	7.0	78	100	0	0	null	null	null
52834	tt0068653	Gran Canaria	null	null	75	0	0	null	null	null

Figura 47. Tabla hechos contenidoAudiovisual (parte 1) - Elaboración propia

id_guionista	id_director	id_actorprincipal	id_actorsecundario1	id_actorsecundario2	id_tipo	id_produccion	id_plataforma	id_premiomasimportante	id_idioma	id_generopri
65003	122	9	182	893	1	221	3	21	1	8
44486	38772	106116	182017	110568	1	194	1	1	214	10
485	522	25591	56576	746	1	221	1	1	72	7
20458	35374	99993	61734	109848	1	221	1	1	72	7
44853	40943	94193	11979	12659	1	106	1	1	118	2
4429	25677	776030	776028	776029	1	106	1	1	118	10
6229	4949	113195	178114	119742	1	228	1	1	88	10
2478	2377	122867	182026	139241	1	195	1	1	241	16

Figura 48. Tabla hechos contenidoAudiovisual (parte 2) - Elaboración propia

6.3.2.13 Carga tablas Hive

Antes continuación, tendremos que cargar (load) las tablas del almacén de datos que tenemos en Hive, al entorno SQL server para posteriormente poder realizar el cubo OLAP, para ello lo primero que hay que hacer es configurar el conector ODBC añadiéndole los parámetros de la máquina guest que es donde está nuestro almacén de datos.

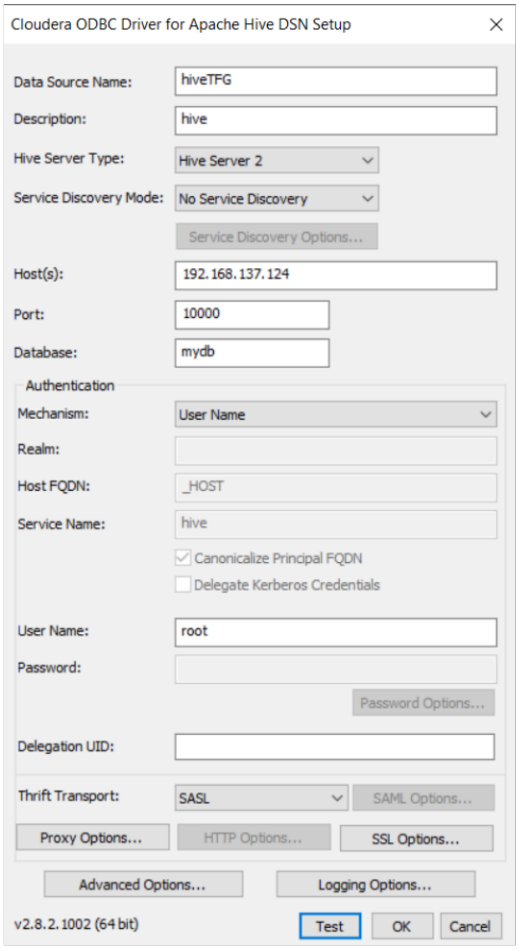


Figura 49. Creación conector ODBC – Elaboración propia

Posteriormente, en SSMS creamos una conexión a la base de datos, indicando como nombre del servidor localhost y usando Windows Auth, para que no haga falta introducir ningún usuario ni contraseña.

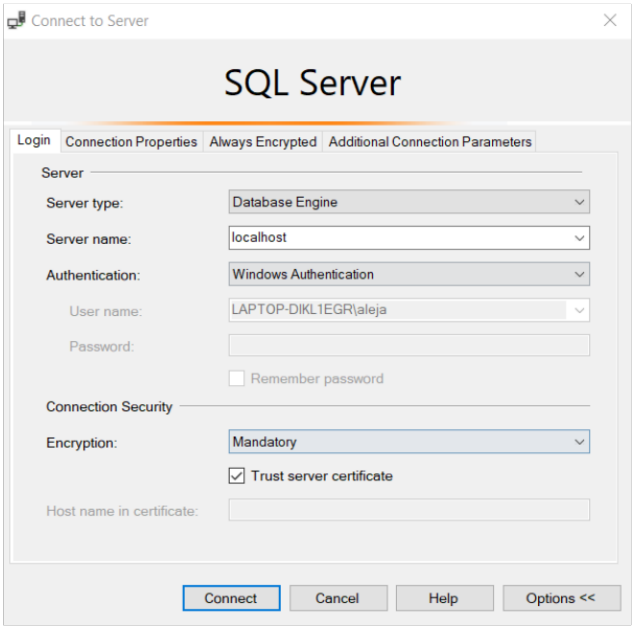


Figura 50. Creación database SQL – Elaboración propia

Posteriormente, en la sección "Linked Objects" ponemos los parámetros de nuestro ODBC y cargamos las tablas del almacén de datos.

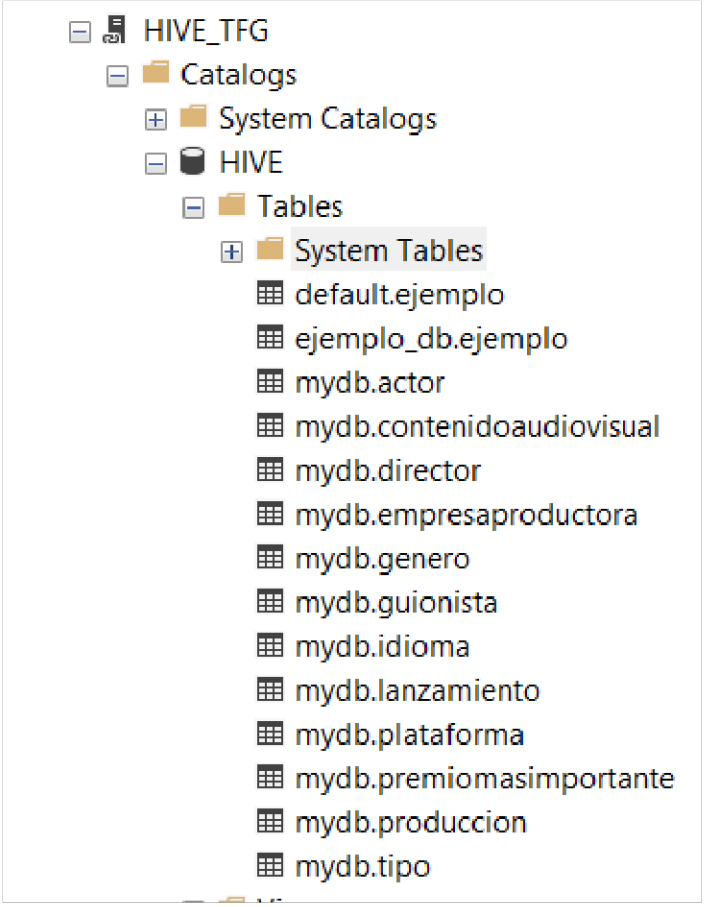


Figura 51. Carga de tablas del almacén de datos

Posteriormente definimos la base HiveLinked y, utilizando querys creamos vistas para todas las dimensiones y la tabla de hechos. Esto lo hacemos para poder cargarlas posteriormente en Visual Studio.

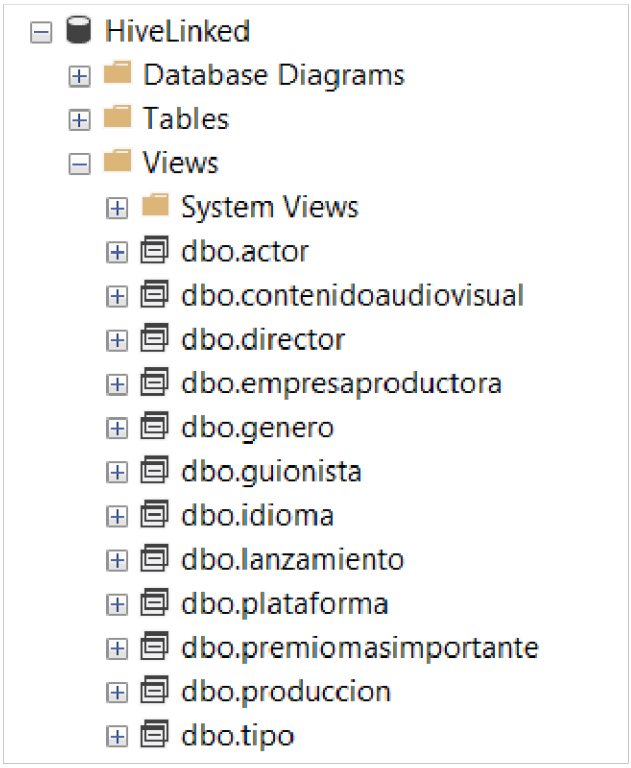


Figura 52. Vistas tablas – Elaboración propia

7. Cubos OLAP

Antes de visualizar los datos en forma de paneles e informes, crearemos un cubo multidimensional el cual nos permitirá realizar análisis OLAP (Online Analytical Proccesing) como consultas. Para ello la herramienta que usaré es SQL Server Management Studio (SSMS) para cargar las diferentes tablas de Hive del almacén de datos, para ello se utilizará también el conector ODBC de Cloudera. Por otro lado, también se utilizará Visual Studio en modo multidimensional y se creará un proyecto de Analysis Services para diseñar y desplegar el cubo.

El precálculo es un aspecto importante en SSMS ya que construye y almacenas agregaciones como sumas, recuentos, medias, de antemano, lo que acelerar las consultas.

7.1 Diseño cubo OLAP

A continuación, se procede al diseño de nuestro cubo, para ello creamos un proyecto multidimensional en Visual Studio. Nos aseguramos de que está apuntando a localhost y hacia la base de datos que hemos creado con las vistas.

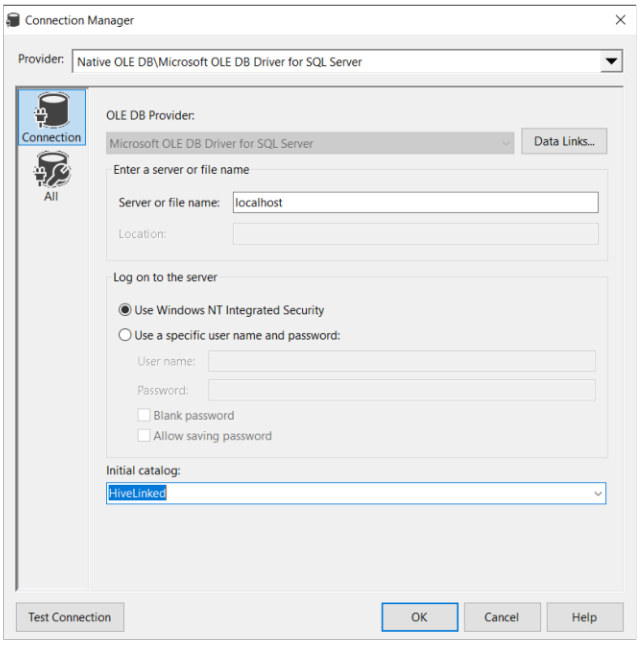


Figura 53. Conexión BD – Elaboración propia

Posteriormente debemos añadir todas las vistas de las diferentes dimensiones y de la tabla de hechos que creamos en HiveLinked.

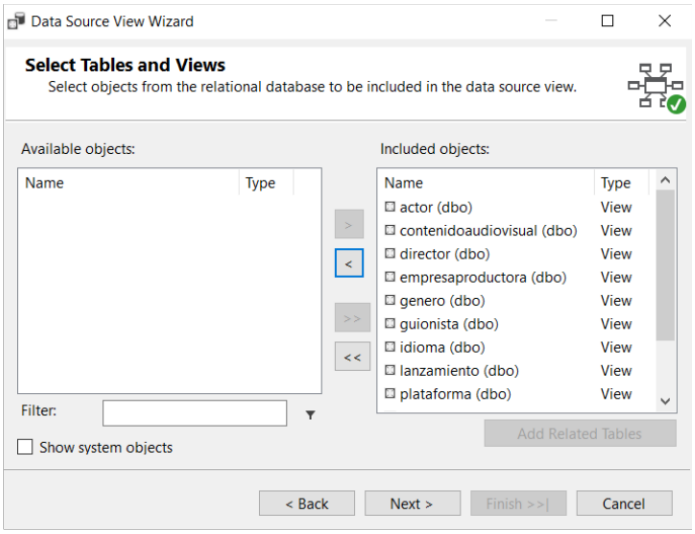


Figura 54. Carga vistas – Elaboración propia

Una vez que hemos cargado todas las vistas, procedemos a crear las diferentes dimensiones. Cada dimensión tendrá un identificador que será su id. También añadimos los atributos que aporten información descriptiva, por ejemplo, en actor seleccionamos el nombre, el país y su reputación.

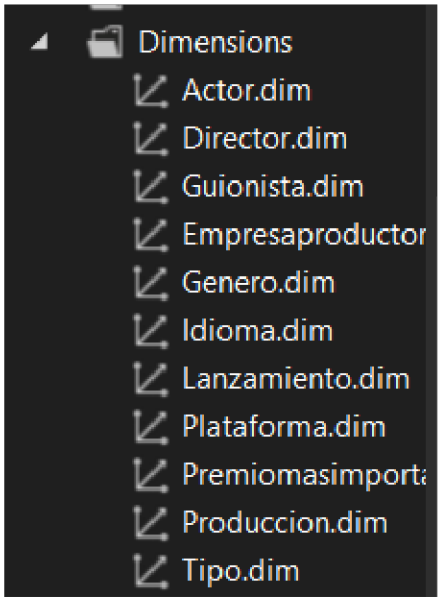


Figura 55. Dimensiones creadas – Elaboración propia

En total tenemos 11 dimensiones además de la tabla de hechos.

Cabe destacar que para la dimensión lanzamiento hay que crear una jerarquía de tiempo, como diseñamos anteriormente en el diseño conceptual que agrupe primero por mes, luego por época y por último por año: mes → época → año. Para ello lo tendremos que crear en la pantalla Hierarchies.

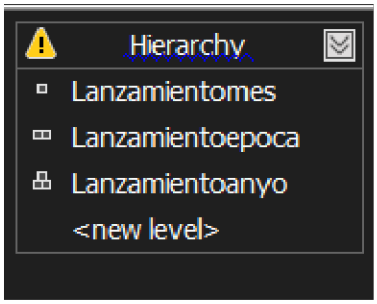


Figura 56. Jerarquía temporal – Elaboración propia

Posteriormente creamos el cubo, para ello seleccionamos la tabla de hechos, y seleccionamos las medidas y las diferentes dimensiones.

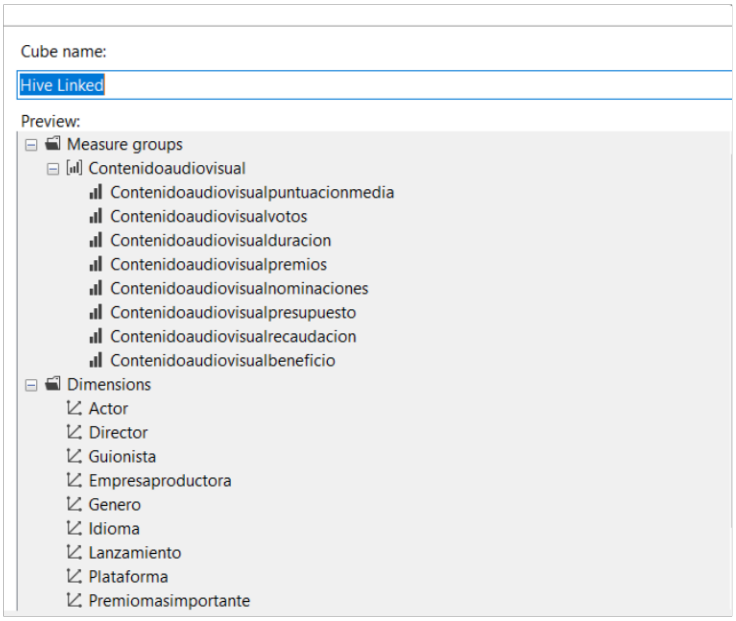


Figura 57. Medidas y dimensiones cubo – Elaboración propia

Una vez que tenemos el cubo creado, establecemos las relaciones entre las dimensiones y la tabla de hechos.

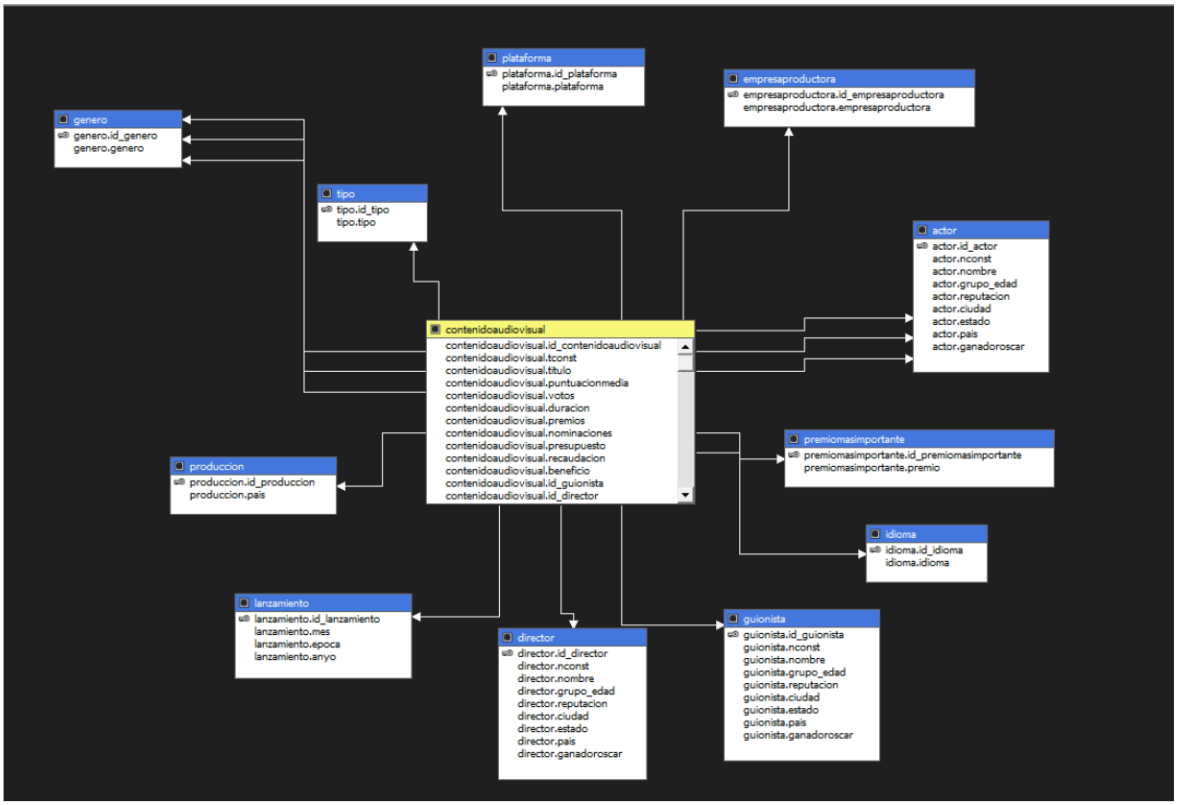


Figura 58. Relaciones dimensiones y tabla de hechos – Elaboración propia

7.2 Despliegue y procesamiento

Una vez que tenemos diseñado el cubo OLAP, procedemos a cargar los datos. Primero cargamos las dimensiones de manera individual y posteriormente la tabla de hechos.

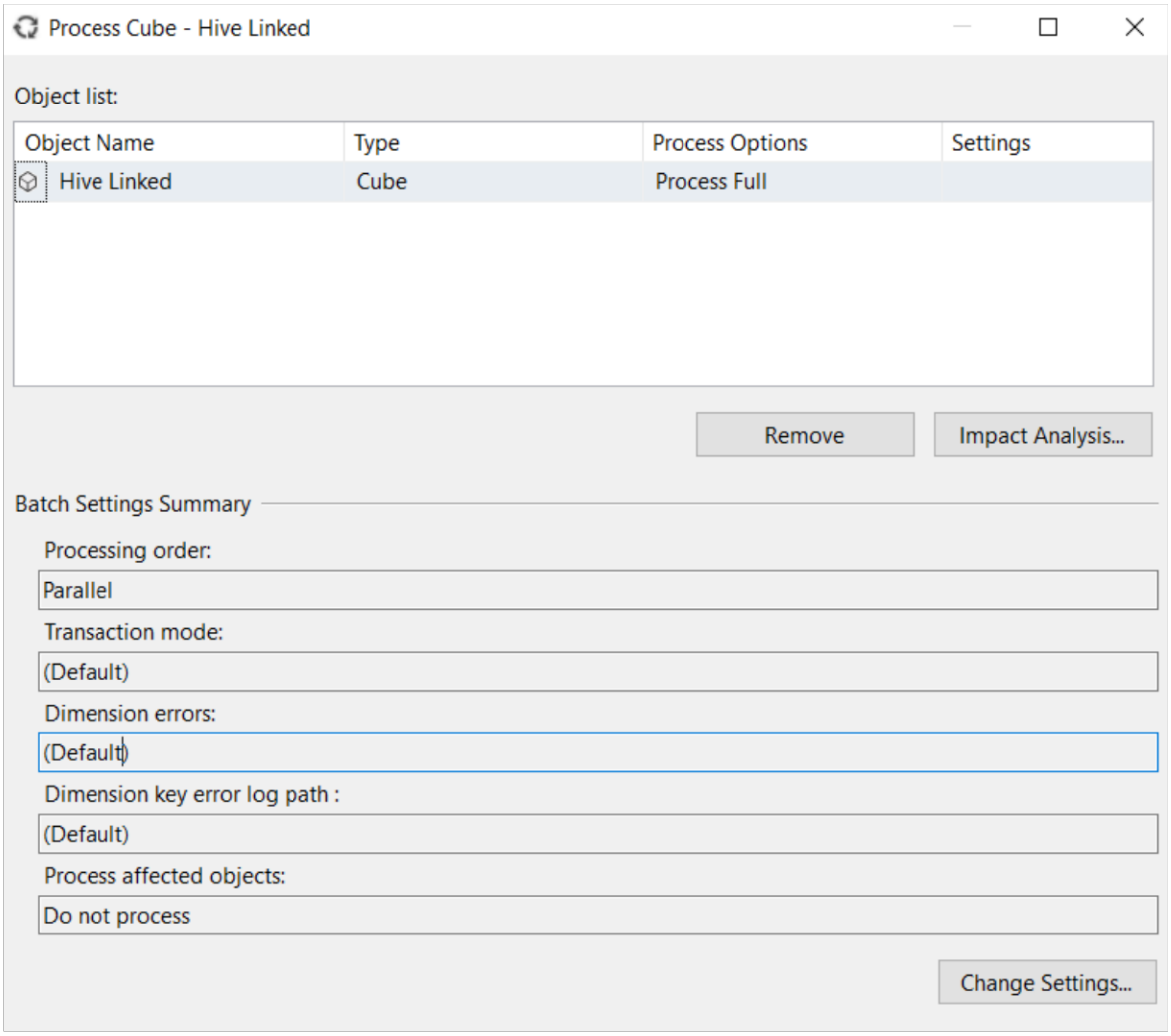


Figura 59. Carga datos cubo – Elaboración propia

Una vez que tenemos todos los datos cargados, podremos explorar el cubo y generar análisis OLAP.

A continuación, vamos a ver los premios que ganan algunos de los diferentes géneros de los contenidos audiovisuales.

Generogenero	Contenidoaudiovisual Count	Contenidoaudiovisualpremios
Action	70210	38000
Adult	11524	621
Adventure	29212	17266
Animation	19260	5085
Biography	24046	23607
Comedy	177524	78802
Crime	35653	27267
Documentary	191388	44016
Drama	238714	151783
Family	16969	2468
Fantasy	7169	2309
Film-Noir	34	3
Game-Show	7215	1039
History	4630	801
Horror	28538	8500
Music	14842	1016

Figura 60. Premios por género – Elaboración propia

Como podemos observar de los géneros mostrados, los contenidos premiados que más premios tienen son los que pertenecen al género Drama con más de 150000 contenidos premiados de aproximadamente 240000 contenidos lo que supone más de la mitad de los, contenidos de Drama están premiados. Por otro lado, los contenidos menos premiados son los de Film-Noir con 3 premios de un total de 34 contenidos.

Sin embargo, el género Documentary que es el segundo que más contenidos tiene de casi 200000 contenidos 44000 han sido premiados es decir un 23% de dichos contenidos tienen premios y tienen menos premios que otros géneros que hacen menor cantidad de contenidos como Comedy.

Entonces, aumentar la producción de contenidos sí que ayuda a que esos géneros tengan más premios, pero no lo garantiza, ya que hay géneros con muchísima producción y su proporción en cuanto a contenidos premiados es bajo como Documentary y otros géneros con menos contenidos que consiguen más premios como Comedy.

En resumen, el cubo OLAP, nos sirve para realizar consultas con millones de datos que en Hive tardarían un tiempo mucho mayor. Además, podremos cargar este cubo en otras herramientas como Power BI, lo que aportará un mayor rendimiento y usabilidad.

8. Cuadros de mandos

La última parte es la visualización de datos, que como dijimos anteriormente, para ello utilizaremos la herramienta Power BI. Esta herramienta, solo está disponible para Windows, y al tener nuestro Data Warehouse en Linux debemos usar un conector. En mi caso voy a utilizar el conector ODBC de Cloudera. Previamente debemos realizar algunas configuraciones en nuestra red para poder cargar los datos de nuestro Data Warehouse.

Una vez que hemos creado el conector, en Power BI lo seleccionamos las tablas que queremos y procedemos a cargar los datos

A continuación, cargamos las tablas de nuestro almacén de datos en Power BI y procedemos a definir las relaciones entre claves ajenas.

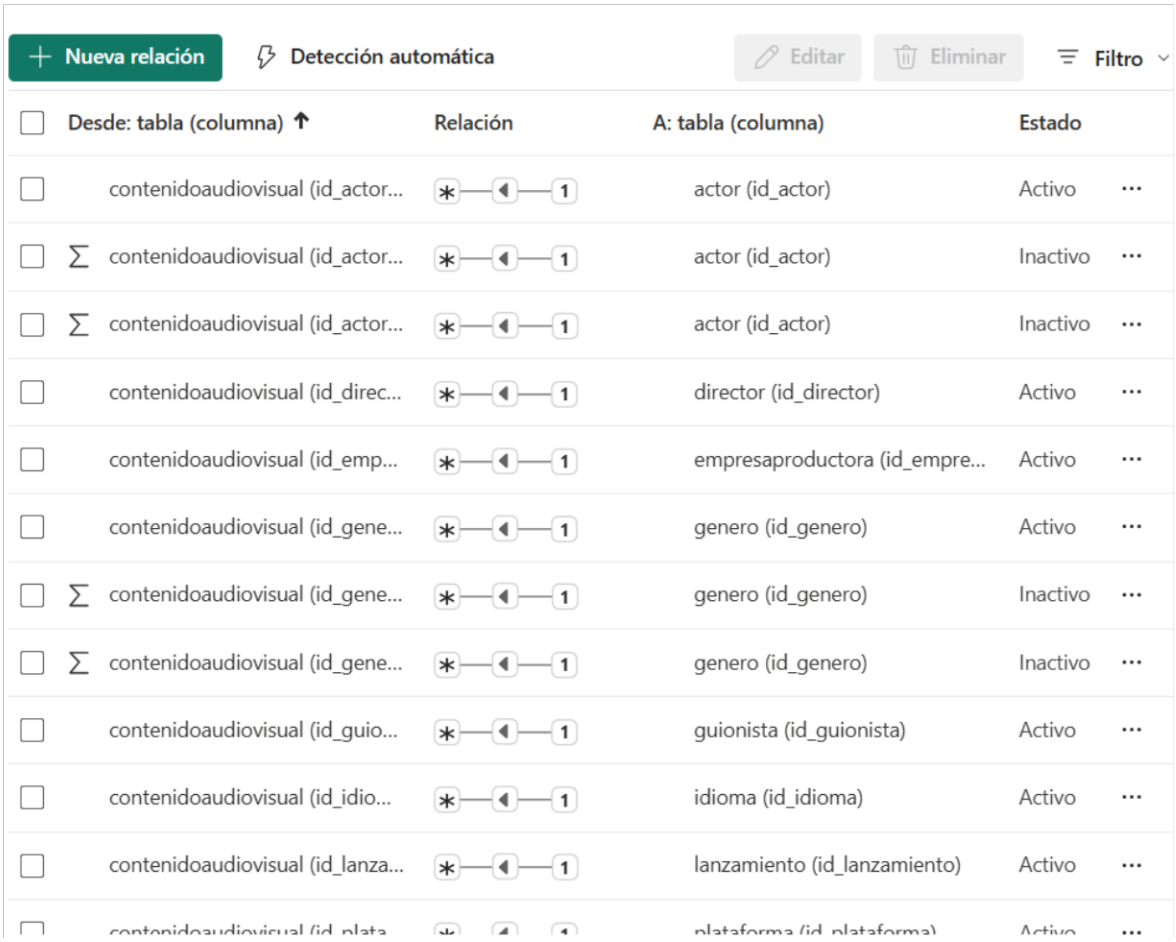


Figura 61. Relaciones claves ajenas Power BI – Elaboración propia.

8.1 Dashboard, objetivos y KPIs

A continuación, podemos observar el dashboard que se ha diseñado para poder responder a algunas preguntas analíticas que plantearemos más adelante y además permitirá evaluar el grado de cumplimiento de los objetivos y KPIs que hemos planteado en el apartado 3. En él, podemos ver para cada objetivo, la visualización que contiene la información necesaria para responder a los objetivos planteados junto a su KPI.

El dashboard se divide en varias páginas. A continuación, podemos observar la pantalla principal del dashboard, la cual contiene varias herramientas que permite al usuario filtrar la información de los contenidos por diferentes variables como año, país y género. En esta página podemos ver la información más relevante sobre todos los contenidos audiovisuales que hay en el almacén de datos y los contenidos que analizaremos son desde 1894 hasta 2024.

Figura 62. Página principal del dashboard – Elaboración propia

Como podemos ver en la página principal del dashboard tenemos un total aproximado de 1 millón de contenidos, que tienen una puntuación media de 6,37. De dichos contenidos más del 75% son películas y el 25 % son series.

Además, tenemos un gráfico de barras en donde aparecen la cantidad de contenidos que tienen diferentes plataformas, en donde podemos ver que Prime Video es la que más contenido tiene con más de 1000 películas y más de 200 series, seguido muy de cerca por Netflix que cuenta con más de 800 películas y más de 500 series. Como podemos observar Netflix apuesta más por las series que Prime Video. Posteriormente, encontramos los contenidos Multiplataforma, es decir que pertenecen a más de una plataforma, seguido de

AppleTV+, que apuesta más por las series que por las películas, como podemos observar hay una gran diferencia de contenidos con estas 2 últimas con respecto a Netflix y Prime Video.

A continuación, vemos un gráfico de barras en donde aparecen el top de 5 géneros de contenidos. Como podemos ver los géneros que más destacan son "Drama", "Documentary", "Comedy", "Action" y "Crime".

Posteriormente, vemos un gráfico de líneas con las diferentes épocas de lanzamientos de los contenidos, en donde a partir del año 2000 vemos un crecimiento masivo en cuanto a la producción de contenidos. Además, podemos ver que históricamente, la época del año que más contenidos se han realizado ha sido en Otoño.

Por último, vemos dos gráficos circulares, el primero que nos muestra el top de los 10 países en donde se han hecho más contenido audiovisuales, en donde podemos ver que destaca Estados Unidos con el 44% de los contenidos, seguido de Reino Unido con un 11% de contenidos y otros países más.

Por otro lado, el otro gráfico nos muestra el top 10 que están realizados los contenidos, en donde podemos observar que el 52% su lenguaje original es el inglés. Luego tenemos un 13% de contenidos que los tenemos con multilenguaje varios idiomas, seguido del español con un 7% y otros idiomas más.

Esta pantalla principal nos ayudará a responder a las preguntas analíticas que se plantearán más adelante tanto de contenidos actuales como otros más antiguos.

A continuación, se proponen los objetivos con sus visualizaciones correspondientes.

8.1.1 Análisis de las producciones de cine nacional

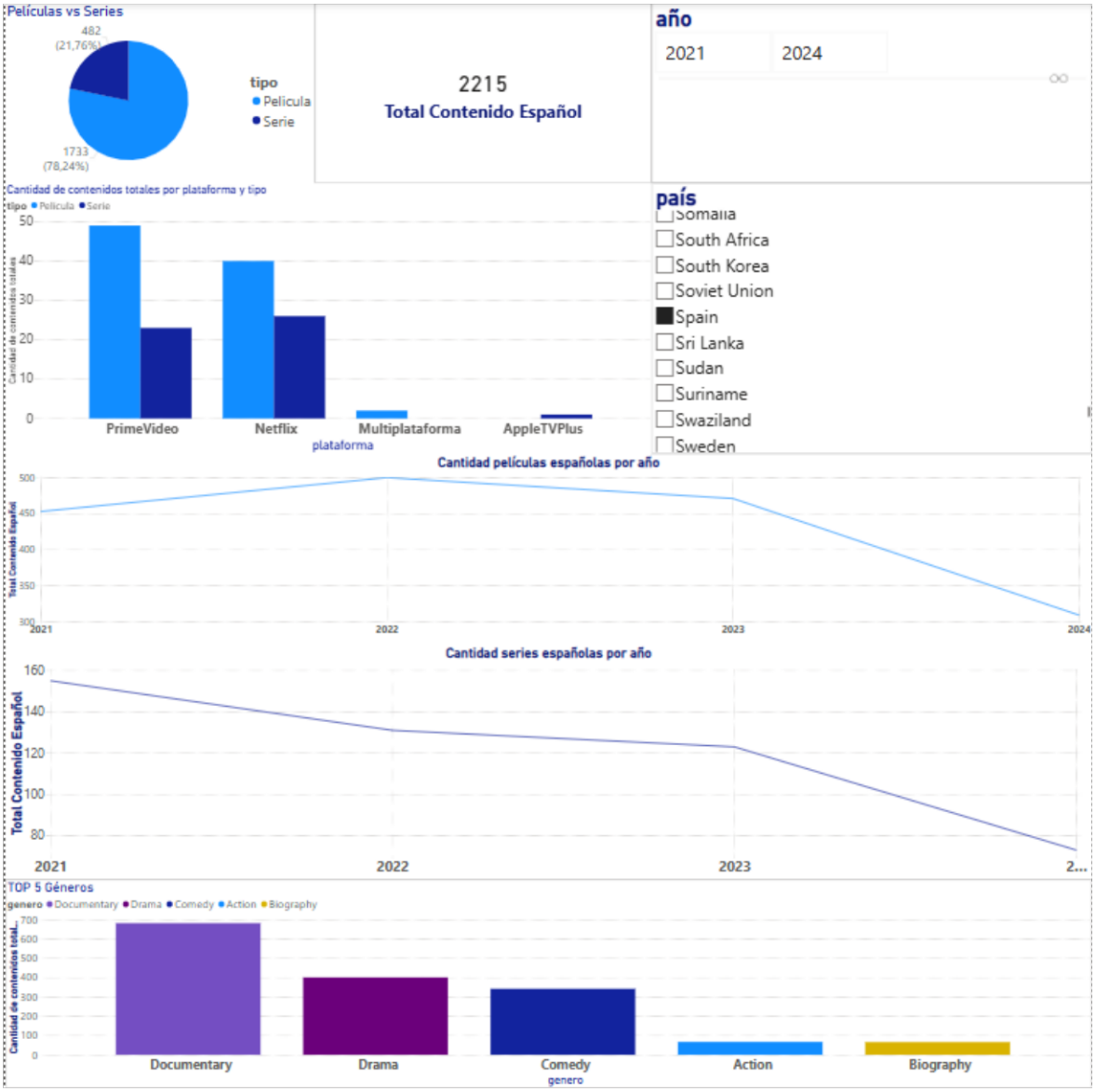


Figura 63. Visualización contenidos nacionales – Elaboración propia

Esta es la página del dashboard que se ha realizado para analizar las producciones de cine nacional. En ella se puede ver que se ha incluido una parte de filtros para filtrar la información por contenidos nacionales y un filtro para seleccionar la información en un periodo de tiempo concreto. Además, hay un par de gráficos de líneas donde permite analizar la evolución temporal de los contenidos, así como otros aspectos como el tipo o la plataforma de los contenidos.

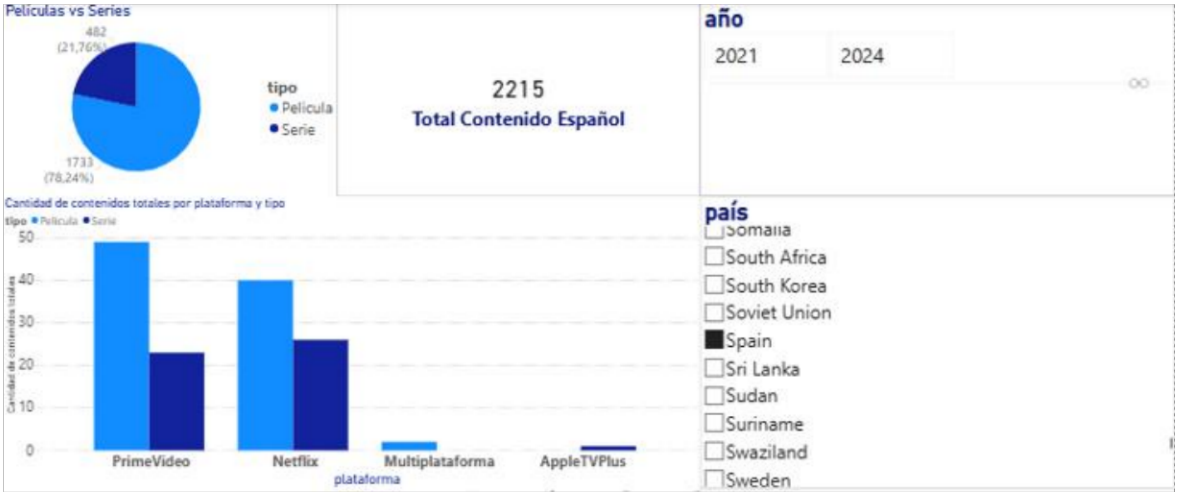


Figura 64. Gráficos tipo y plataformas contenidos nacionales 2021 - 2024 – Elaboración propia

En la imagen anterior podemos observar que en nuestro almacén de datos hay un total de 2215 contenidos españoles que su fecha de lanzamiento se encuentra entre 2021 y 2024. En el gráfico circular podemos observar que de esos contenidos predominan las películas sobre las series nacionales, aproximadamente un 78% de los contenidos nacionales son películas. Mientras que en el gráfico de barras podemos observar que la plataforma que más contenidos nacionales tiene es Prime Video con 49 películas y 23 series, seguido de Netflix con 40 películas y 26 series, la plataforma que menos contenido tiene es AppleTV+ con solamente una serie nacional.

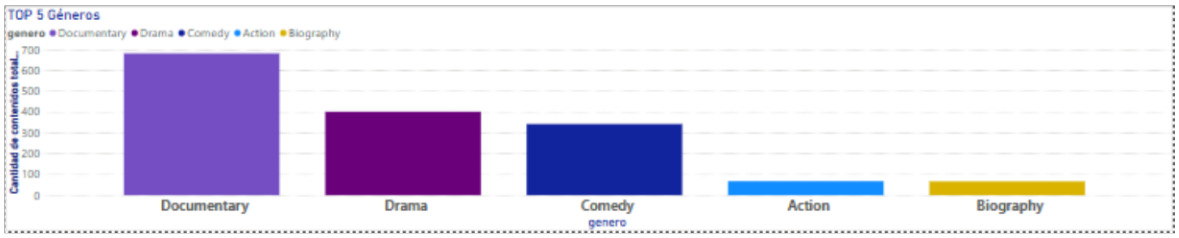


Figura 65. Gráfico géneros nacionales más producidos 2021-2024 – Elaboración propia

En el gráfico de barras anterior podemos ver los géneros que más se han producido en España durante 2021 y 2024. Podemos observar que el género al que más contenidos pertenecen con una diferencia importante con el resto es "Documentary" con casi 700 contenidos, lo sigue el género "Drama" con aproximadamente 400 contenidos y en tercer lugar está el género "Comedy" con 340 contenidos.

Se pretende analizar si se cumple el objetivo nacional del Plan Fomento del cine español 2021 – 2024 del ICAA (Instituto de la Cinematografía y de las Artes Audiovisuales), el cual se marcó como objetivo un aumento del 30% de los contenidos 2021 a 2024 tanto en películas como en series, con el fin de aumentar la presencia de contenidos nacionales después de la crisis del COVID.

Para este objetivo a continuación se definen dos KPIs, uno para comprobar si hubo un aumento de películas nacionales y otro para series nacionales.

8.1.1.1 KPI 1.1: Comprobar si el crecimiento de películas españolas aumenta en un +30% del 2021 al 2024.

Figura 66. Gráfico evolución películas nacionales 2021-2024 – Elaboración propia

Para ver si se cumple el KPI 1.1 analizaremos el gráfico de líneas anterior que muestra la cantidad de películas españolas que se han producido anualmente durante los años 2021 y 2024. Como podemos observar, en los últimos años no solo no ha aumentado la producción de películas nacionales, sino que han disminuido siendo el año con mayor caída en 2024.

A continuación, podemos ver la tabla de variación anual con respecto al año anterior para ver cuanto aumenta o disminuyen las películas nacionales y ver más detalladamente esa caída general de películas nacionales.

Año	Suma de Datos	Suma de Variación anual
2021	453	
2022	500	0,10
2023	471	-0,06
2024	309	-0,34
Total	**1733**	**-0,30**

Figura 67. Evolución anual de producción de películas españolas – Elaboración propia

En la tabla anterior podemos apreciar cómo varían año a año la producción de películas nacionales. Podemos observar que en 2022 se produce un aumento del +10% en producción de películas nacionales con respecto al año anterior en donde es el máximo aumento, mientras que la mayor caída en cuanto a producción de películas se produce en 2024 con un -36% con respecto al año anterior.

Como podemos observar en los datos, no solo **no se cumple** el KPI 1.1 ya que no solo no se aumentan la producción de películas +30%, sino que por el contrario las películas han disminuido un -30% su producción desde 2021 a 2024.

8.1.1.2 KPI 1.2: Comprobar si el crecimiento de series españolas aumenta en un +30% del 2021 al 2024.

Figura 68. Gráfico evolución series nacionales 2021-2024 – Elaboración propia

Para ver si se cumple el KPI 1.2 analizaremos el gráfico de líneas anterior que muestra la cantidad de series españolas que se han producido durante los años 2021 y 2024. Como podemos observar, en los últimos años no solo no ha aumentado la producción de series nacionales, sino que por el contrario cada año ha ido disminuyendo la producción de series con respecto al año anterior, siendo 2024 el año que ha habido una mayor caída en la producción de series nacionales.

En la siguiente tabla podemos ver la variación anual con respecto al año anterior para ver el decremento de las series nacionales y ver más detalladamente la caída general en la producción de series nacionales desde 2021 a 2024.

Año	Suma de Datos	Suma de Variación anual
2021	155	
2022	131	-0,15
2023	123	-0,06
2024	73	-0,41
Total	**482**	**-0,62**

Figura 69. Evolución anual de producciones de españolas – Elaboración propia

Como podemos observar en la tabla anterior, las series españolas han sufrido una caída muy notable ya que desde 2021 solo han disminuido su producción en donde destaca el año 2024 con un -41% en la reducción de series españolas con respecto al año anterior.

Las series españolas han disminuido su producción un -62% del 2021 al 2024, por lo que **no se cumple** el KPI 1.2 que buscaba aumentar las series españolas un +30% en ese periodo.

En resumen, ninguno de los objetivos definido por el ICAA (Instituto de la Cinematografía y de las Artes Audiovisuales) se cumpliría en nuestro almacén de datos.

8.1.2 Análisis de las producciones más populares

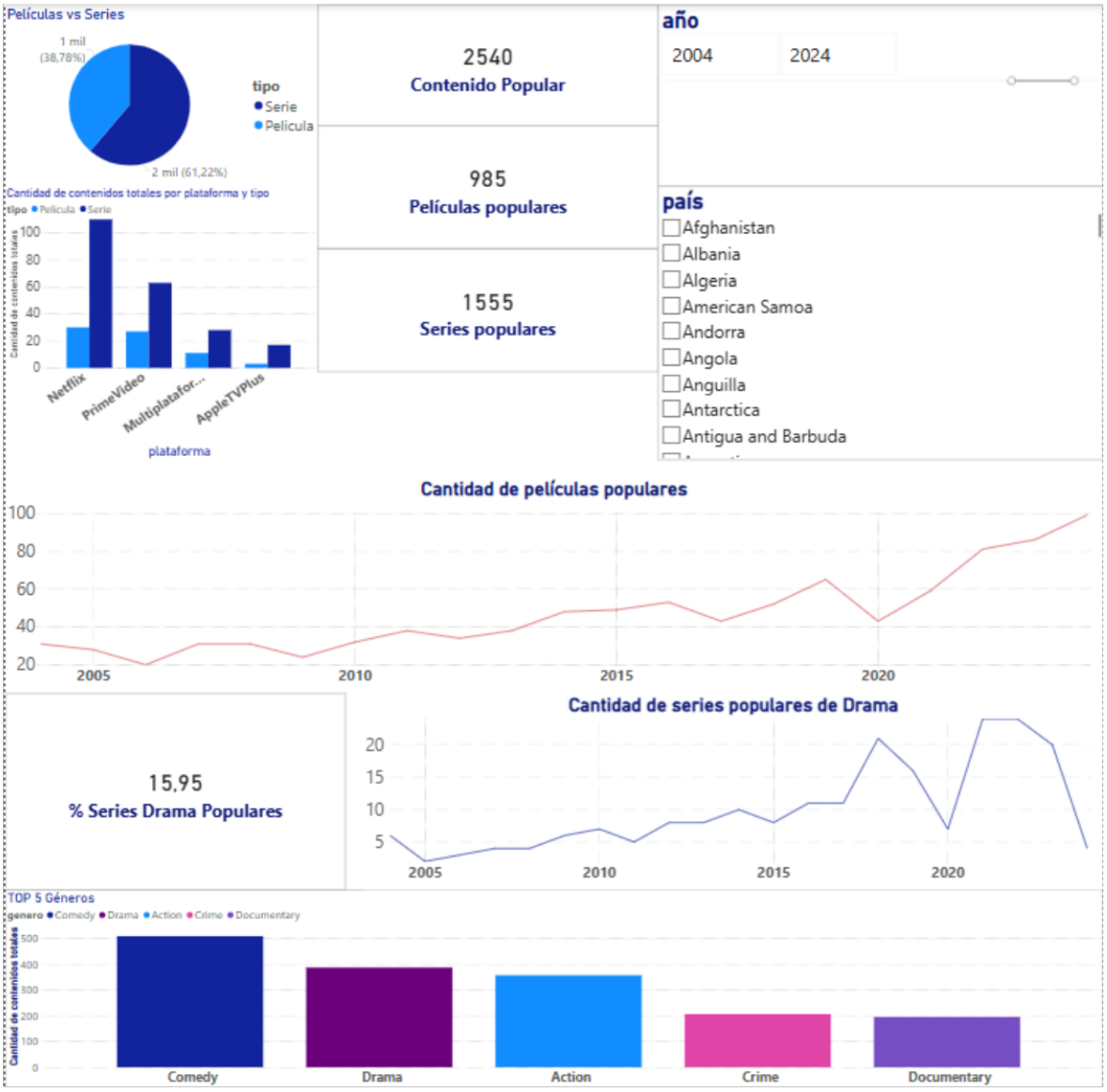


Figura 70. Visualización contenidos populares – Elaboración propia.

La imagen anterior muestra la página del dashboard que se ha realizado para analizar las producciones más populares. En ella se puede ver que se ha incluido una parte de filtros por si el usuario desea filtrar la información de los contenidos por un país concreto o un periodo de tiempo concreto.

Además, hay un par de gráficos de líneas donde permite analizar la evolución temporal de la producción de contenidos populares, así como otros aspectos como el tipo o la plataforma de los contenidos.

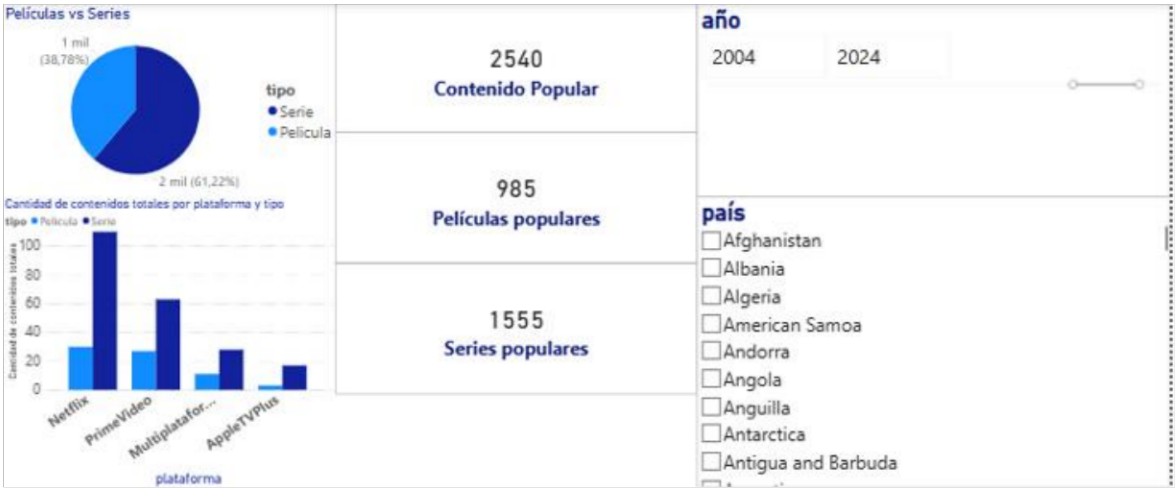


Figura 71. Gráficos tipo y plataformas contenidos populares 2004 - 2024 – Elaboración propia

En la imagen anterior podemos observar que en nuestro almacén de datos hay un total de 2540 contenidos populares de 2004 a 2024. En el gráfico circular observamos que hay más series populares que películas, aproximadamente el 61% de contenidos populares, hay 1555 series populares, mientras que el 39% restante son películas populares es decir hay 985 películas populares. Por otro lado, en el gráfico de barras podemos observar que la plataforma que más contenidos populares tiene es Netflix con 110 series y 30 películas, lo sigue Prime Video 63 películas y más de 27 películas, mientras que la plataforma que menos contenido popular tiene es AppleTV+ con 17 series y 3 películas.

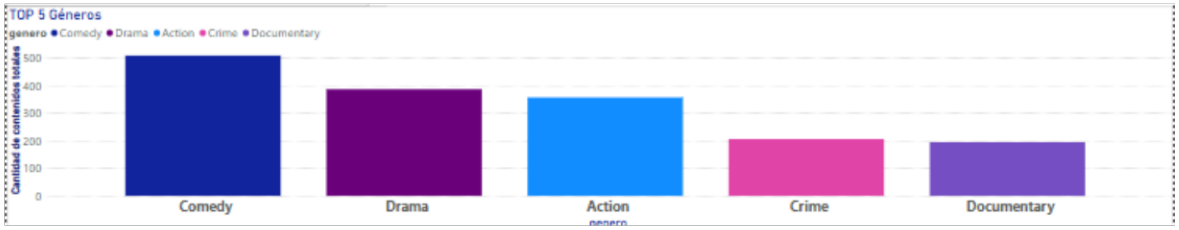


Figura 72. Gráfico géneros contenidos más populares 2004-2024 – Elaboración propia

En el gráfico de barras anterior podemos ver los géneros a los que pertenecen los contenidos más populares durante 2004 y 2024. Se puede observar que el género al que pertenecen más contenidos populares es el género "Comedy" con más de 500 contenidos, el segundo género con más contenidos populares es "Drama" con 387 contenidos, lo sigue el género "Action" con 357 contenidos populares.

Se pretende analizar cuál es la tendencia de la producción de los contenidos más populares en los últimos 20 años. Los contenidos en IMDb están puntuados de 0 a 10 por lo que analizaremos aquellos contenidos que tienen al menos una puntuación de 8. También es importante destacar que un contenido puede tener una puntuación de 10 y por ejemplo 5 votos, por los contenidos analizados también deben tener al menos 1000 votos.

Para este objetivo a continuación se definen dos KPIs, uno para comprobar que no hubo una bajada en la producción de películas populares y otro para ver si al menos un 10% de las series populares son del género "Drama" en los últimos 20 años [2004 – 2024].

8.1.2.1 KPI 2.1: Analizar si las películas populares no han bajado su producción en un -15% en los últimos 20 años [2004 – 2024]

Figura 73. Gráfico evolución películas populares 2004-2024 – Elaboración propia

Para ver si se cumple el KPI 2.1 analizaremos el gráfico de líneas anterior que muestra la cantidad de películas populares que se han producido de manera anual durante los últimos 20 años [2004 – 2024]. Como podemos observar, hay un aumento general de las películas populares en los últimos 20 años, en donde curiosamente entre los años 2020 y 2022 vemos un aumento notable en la producción de películas populares.

A continuación, podemos ver la tabla de variación anual con respecto al año anterior para ver cuanto ese aumento general de manera más detallada.

Año	Suma de Cantidad	Suma de Variación anual
2004	31	
2005	28	-0,10
2006	20	-0,29
2007	31	0,55
2008	31	0,00
2009	24	-0,23
2010	32	0,33
2011	38	0,19
2012	34	-0,11
2013	38	0,12
2014	48	0,26
2015	49	0,02
2016	53	0,08
2017	43	-0,19
2018	52	0,21
2019	65	0,25
2020	43	-0,34
2021	59	0,37
2022	81	0,37
2023	86	0,06
2024	99	0,15
Total	**985**	**1,73**

Figura 74. Evolución producciones contenidos populares – Elaboración propia.

Como podemos ver en la tabla anterior, la producción de películas populares aumenta de forma notable en los últimos años, llegando a alcanzar un aumento total del +173% en los últimos 20 años.

El año que más crecimiento de producción de películas populares hubo con respecto al año anterior fue en 2007 con un aumento en la producción de +55%, mientras que la mayor caída en la producción fue en 2006 con un -29%.

La tendencia es muy positiva y **cumple ampliamente** el KPI 2.1 el cual no permite una bajada en las producciones de películas populares un -15%, de hecho, aumenta en un +173% el total de producciones populares en los últimos 20 años.

8.1.2.2 KPI 2.2: Analizar si pertenecen al género "Drama" al menos un 10% de las series populares en los últimos 20 años [2004 – 2024]

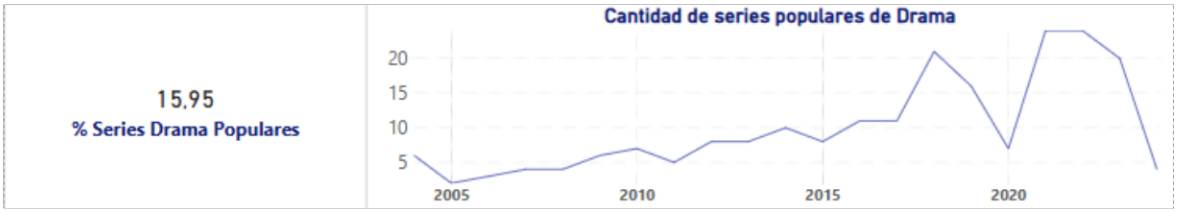


Figura 75. Gráfico evolución serie dramas populares y porcentaje 2004-2024 – Elaboración propia

Para ver si se cumple el KPI 2.2 analizaremos la imagen anterior la cual contiene una tarjeta con el porcentaje de las series de género "Drama" populares, este porcentaje se calculó dividiendo las series de género "Drama" populares entre el total de series populares y dejándolo en formato porcentaje.

Por otro lado, el gráfico de líneas muestra la cantidad de series populares que se han ido realizando durante los últimos 20 años [2004 – 2024]. En el gráfico se puede observar que hay una tendencia ascendente en la producción de series populares de género "Drama", en donde se alcanza el máximo en 2021 y 2022 con 24 series producidas ambos años y por otra parte la mayor caída en la producción de series populares de "Drama" se produce en 2024 ya que pasa de 20 series producidas el año anterior a 4.

Con los datos podemos ver que **se cumple** el KPI 2.2 de que al menos un 10% de las series populares de los últimos 20 años pertenecen al género Drama, siendo este género el segundo género que más producciones se han realizado durante este periodo.

8.1.3 Análisis de los lanzamientos en fechas clave de España y EE. UU

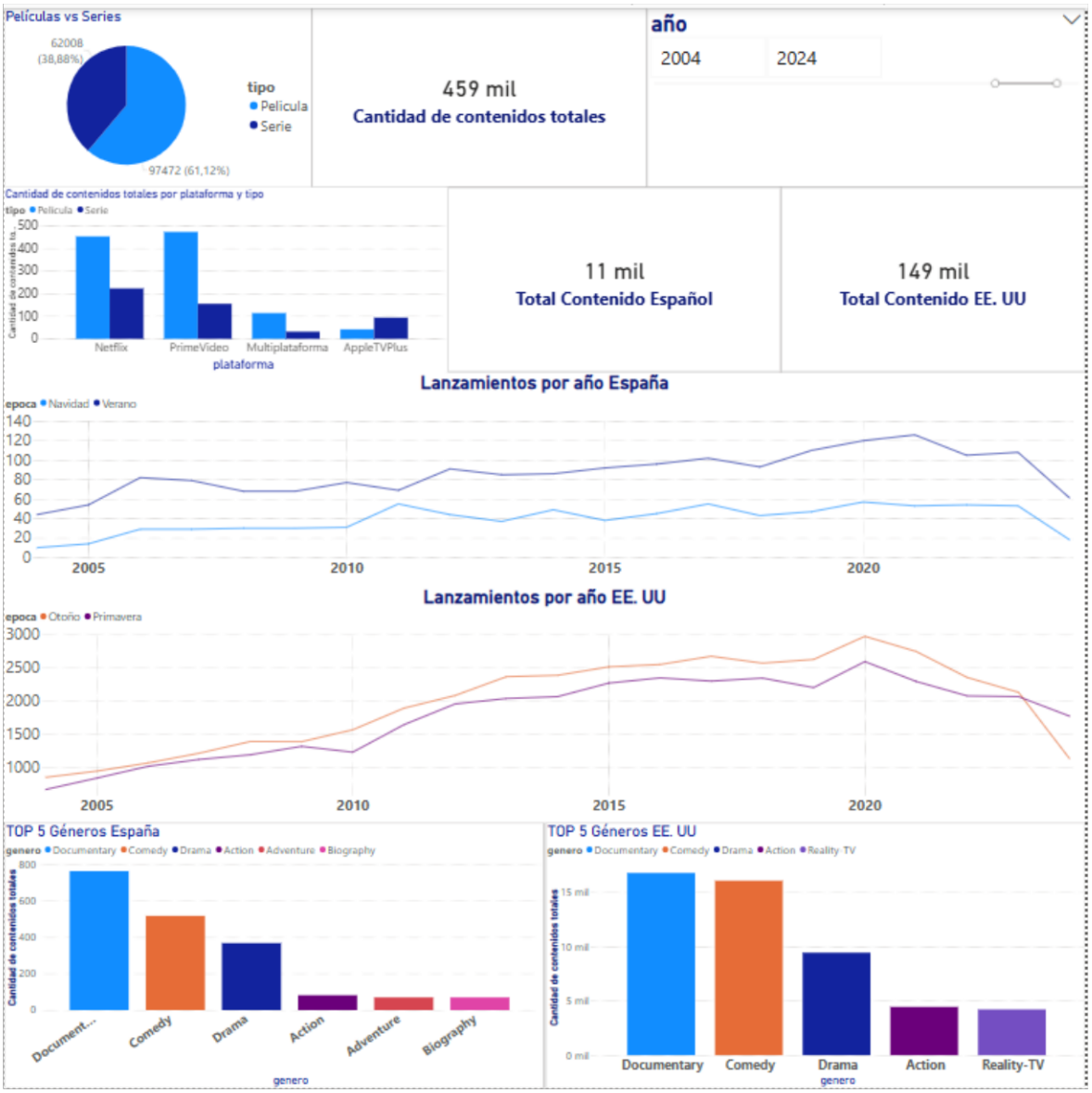


Figura 76. Visualización lanzamiento de contenidos en España y EE. UU – Elaboración propia.

La imagen anterior muestra la página del dashboard que se ha realizado para analizar los contenidos que son lanzados en fechas clave de España y EE. UU. En ella se puede ver que se ha incluido un filtro para que el usuario pueda filtrar la información de los contenidos en un periodo de tiempo concreto.

Además, hay un par de gráficos de líneas donde permite analizar la evolución temporal de la producción de contenidos en fechas clave de España y EE. UU, así como otros aspectos

como los géneros a los que más pertenecen los contenidos lanzados tanto en España como en EE. UU en fechas clave.

Figura 77. Gráfico tipo y cantidad contenidos España y EE. UU 2004 - 2024 – Elaboración propia

En la imagen anterior podemos observar que en nuestro almacén de datos hay un total aproximado de 459000 contenidos de 2004 a 2024, de los cuales 11000 son contenidos de España y 149000 son contenidos de EE. UU. En el gráfico circular observamos que hay más películas que series, aproximadamente el 67% de contenidos son películas y el 33% restante son series.

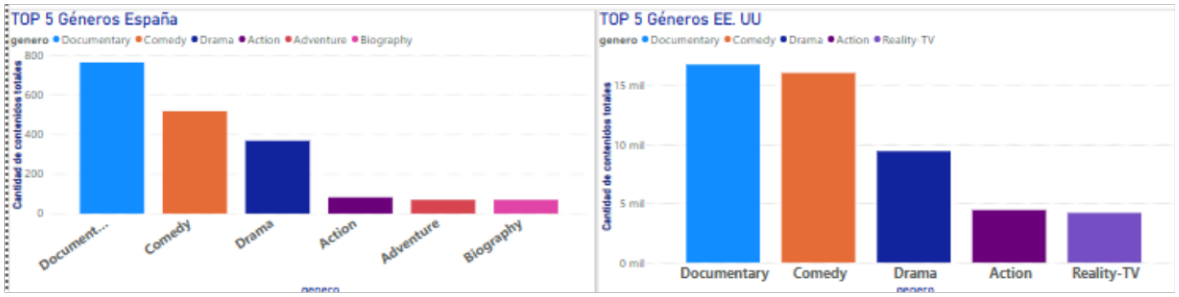


Figura 78. Gráfico géneros con más contenidos en fechas clave en España y EE. UU 2004-2024 – Elaboración propia

En los gráficos de barras anteriores podemos ver los géneros que tienen más contenidos lanzados en fechas clave en España y en EE. UU. Curiosamente, los 4 géneros a los cuales los contenidos lanzados en fecha clave tanto en España como en Estados Unidos coincide siendo los géneros "Documentary", "Comedy", "Drama" y "Action" los más producidos en fechas clave en ambos países.

A continuación, se pretende analizar si en los últimos 20 años ha habido un aumento sostenido en los estrenos en fechas clave tanto en España como en EE. UU. Como se explicó anteriormente las fechas de estreno que están en el almacén son las fechas de lanzamiento en España.

Para este objetivo, se plantean varios KPIs a continuación.

8.1.3.1 KPI 3.1: Analizar que los contenidos de España lanzados en "Navidad" aumentan +15% en los últimos 20 años [2004 – 2024]

8.1.3.2 KPI 3.2: Analizar que los contenidos de España lanzados en "Verano" aumentan +20% en los últimos 20 años [2004 – 2024]

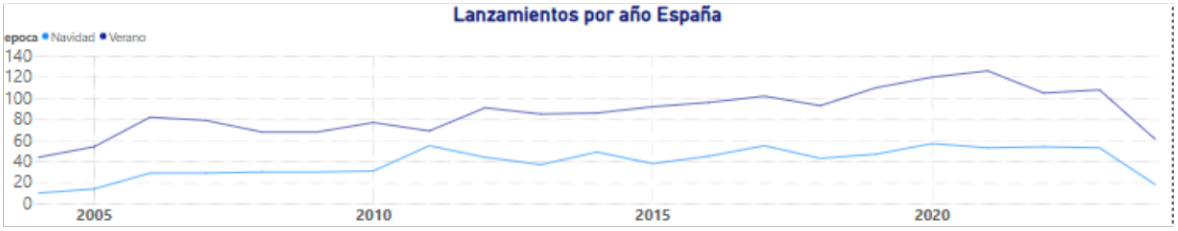


Figura 79. Gráfico lanzamiento contenidos en Navidad y Verano en España 2004-2024 – Elaboración propia

Para conocer si se cumplen los KPI 3.1 y 3.2, analizaremos el gráfico de líneas anterior, en el cual en la línea azul oscura podemos ver la cantidad de contenidos lanzados anualmente en "Verano" y en la línea azul claro podemos ver la cantidad de contenidos lanzados anualmente en "Navidad". Por lo general en ambas épocas del año se puede ver que ha habido un aumento general en la producción de contenidos en España en esas épocas desde 2004 hasta 2024. A continuación, analizaremos más en detalle ese aumento y comprobaremos si se cumplen los KPIs con la tabla de variación.

Año	Suma de Navidad_ES	VarNavidad_ES	Suma de Verano_ES	VarVerano_ES
2004	10		44	
2005	14	0,40	54	0,23
2006	29	1,07	82	0,52
2007	29	0,00	79	-0,04
2008	30	0,03	68	-0,14
2009	30	0,00	68	0,00
2010	31	0,03	77	0,13
2011	55	0,77	69	-0,10
2012	44	-0,20	91	0,32
2013	37	-0,16	85	-0,07
2014	49	0,32	86	0,01
2015	38	-0,22	92	0,07
2016	45	0,18	96	0,04
2017	55	0,22	102	0,06
2018	43	-0,22	93	-0,09
2019	47	0,09	110	0,18
2020	57	0,21	120	0,09
2021	53	-0,07	126	0,05
2022	54	0,02	105	-0,17
2023	53	-0,02	108	0,03
2024	18	-0,66	61	-0,44
Total	**821**	**1,82**	**1816**	**0,70**

Figura 80. Evolución producciones de España en Navidad y Verano – Elaboración propia.

En la tabla anterior, podemos observar que en los últimos 20 años los contenidos españoles que han sido lanzados durante la época de "Navidad" han experimentado un aumento total en su producción del +182% por lo que **sí que cumple** con creces el KPI 3.1 definido que pretende ver si aumenta la producción de contenidos en esta época un +15%.

El año que más contenidos se han producido en "Navidad" es 2020 con 57 contenidos, y siendo 2006 el año que mayor aumento de contenidos hubo con respecto al año anterior con un +107% de aumento en la producción de contenidos. Mientras que por otra parte el año que menos producciones españolas se lanzaron fue en 2004 con 10 contenidos y registrando en 2024 la caída más importante con -66% en la producción de contenidos con respecto al año anterior.

Por otro lado, en cuanto a los contenidos españoles lanzados en "Verano" podemos observar que hay una tendencia al alza registrando un aumento total en la producción del +70%, lo

que **sí que cumpliría** ampliamente el KPI 3.2 que busca comprobar si los contenidos españoles lanzados en esta época del año aumentan al menos un +20% de 2004 hasta 2024. El año en el que se han lanzado más contenidos en "Verano" es 2021 con 126 contenidos, y siendo 2006 el año que mayor aumento de contenidos ha habido con el año anterior con un +52% de aumento en la producción de contenidos. Mientras que por otra parte el año que menos producciones españolas se lanzaron en esta época fue en 2004 con 44 contenidos y registrando en 2024 la caída más importante con -44% en la producción de contenidos con respecto al año anterior.

8.1.3.3 KPI 3.3: Analizar que los contenidos de EE. UU lanzados en "Otoño" aumentan +20% en los últimos 20 años [2004 – 2024].

8.1.3.4 KPI 3.4: Analizar que los contenidos de EE. UU lanzados en "Primavera" aumentan +20% en los últimos 20 años [2004 – 2024].

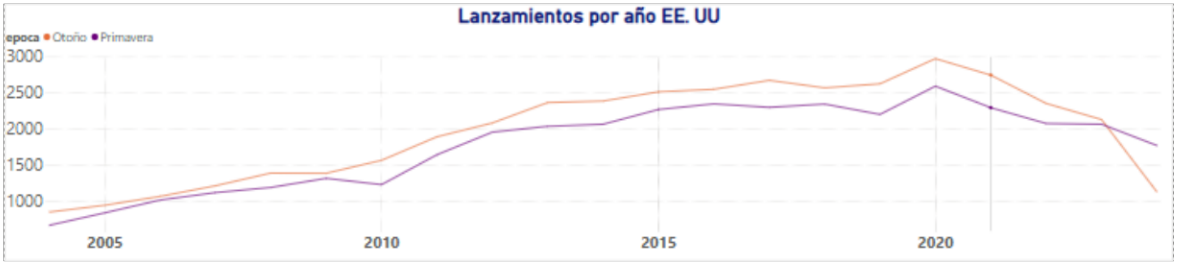


Figura 81. Gráfico lanzamiento contenidos en Otoño y Primavera en EE. UU 2004-2024 – Elaboración propia

Para conocer si se cumplen los KPI 3.3 y 3.4, analizaremos el gráfico de líneas anterior, en el cual en la línea naranja podemos ver la cantidad de contenidos lanzados anualmente en "Otoño" y en la línea morada podemos ver la cantidad de contenidos lanzados anualmente en "Primavera". Por lo general en ambas épocas del año se puede ver que ha habido un aumento general en la producción de contenidos en estadounidenses en esas épocas desde 2004 hasta 2024, no obstante, en 2024 se puede una caída en la producción en ambas épocas del año. A continuación, se realizará un análisis más detallado y comprobaremos si se cumplen los KPIs con la tabla de variación.

Año	Suma de Otono_USA	VarOtono_USA	Suma de Primavera_USA	VarPrimavera_USA
2004	858		675	
2005	952	0,11	847	0,25
2006	1074	0,13	1024	0,21
2007	1221	0,14	1127	0,10
2008	1397	0,14	1198	0,06
2009	1394	0,00	1324	0,11
2010	1573	0,13	1237	-0,07
2011	1898	0,21	1650	0,33
2012	2089	0,10	1962	0,19
2013	2371	0,13	2043	0,04
2014	2392	0,01	2071	0,01
2015	2519	0,05	2275	0,10
2016	2554	0,01	2353	0,03
2017	2678	0,05	2305	-0,02
2018	2574	-0,04	2350	0,02
2019	2630	0,02	2207	-0,06
2020	2978	0,13	2598	0,18
2021	2750	-0,08	2300	-0,11
2022	2359	-0,14	2082	-0,09
2023	2136	-0,09	2072	0,00
2024	1136	-0,47	1777	-0,14
Total	**41533**	**0,55**	**37477**	**1,14**

Figura 82. Evolución producciones de EE. UU en Otoño y Primavera – Elaboración propia.

En cuanto a los contenidos de EE. UU, podemos observar que en los últimos 20 años los contenidos que han sido lanzados durante la época de "Otoño" han experimentado un aumento total en su producción del +55% por lo que **sí que cumple** con creces el KPI 3.3 definido que pretende ver si aumenta la producción de contenidos estadounidenses en esta época un +20%. El año que más contenidos estadounidenses se han producido en "Otoño" es 2020 con 2978 contenidos, y siendo 2011 el año que mayor aumento de contenidos hubo con respecto al año anterior con un +21% de aumento en la producción de contenidos. Mientras que por otra parte el año que menos producciones estadounidenses se lanzaron fue en 2004 con 858 contenidos y registrando en 2024 la caída más importante con -47% en la producción de contenidos con respecto al año anterior.

Por otro lado, en cuanto a los contenidos estadounidenses lanzados en "Primavera" podemos observar que hay una tendencia al alza registrando un aumento total en la producción del

+114%, lo que **sí que cumpliría** ampliamente el KPI 3.4 que busca comprobar si los contenidos estadounidenses lanzados en esta época del año aumentan al menos un +20% de 2004 hasta 2024. El año en el que se han lanzado más contenidos en "Primavera" es 2020 con 2598 contenidos, y siendo 2011 el año que mayor aumento de contenidos ha habido con el año anterior con un +33% de aumento en la producción de contenidos. Mientras que por otra parte el año que menos producciones españolas se lanzaron en esta época fue en 2004 con 675 contenidos y registrando en 2024 la caída más importante con -14% en la producción de contenidos con respecto al año anterior.

Como podemos observar se cumplieron los cuatro KPIs definidos en el objetivo ampliamente.

8.1.4 Análisis de los contenidos de Netflix

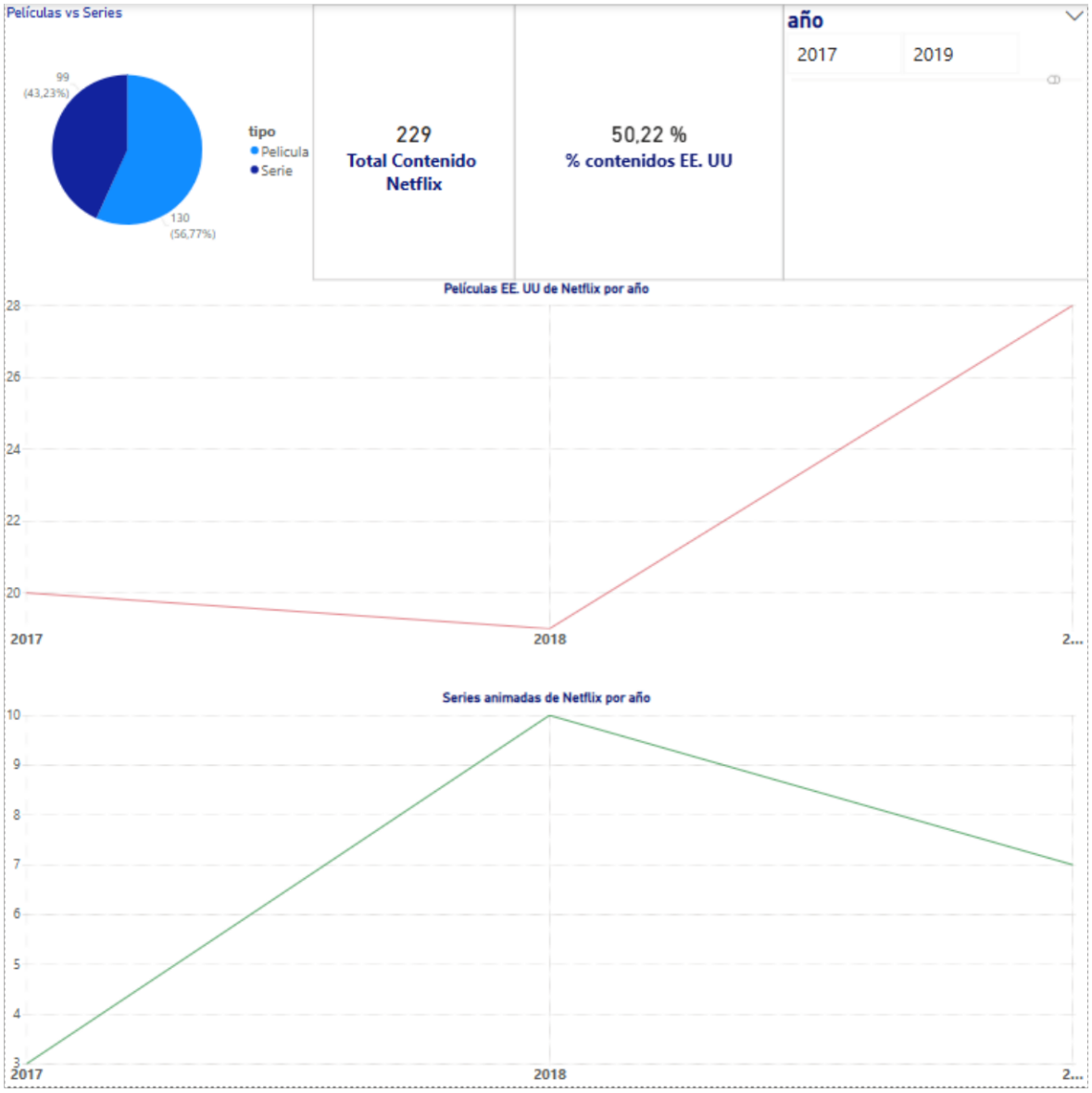


Figura 83. Visualización contenidos Netflix 2017-2019 – Elaboración propia.

La imagen anterior muestra la página del dashboard que se ha realizado para analizar los contenidos de Netflix. En ella se puede ver que se ha incluido un filtro para que el usuario pueda filtrar la información de los contenidos en un periodo de tiempo concreto.

Además, hay un par de gráficos de líneas donde permite analizar la evolución temporal de la producción de contenidos de Netflix.

Se pretende analizar si se cumplen los objetivos marcados por un informe que publicó Netflix del 2017 en el cual se plantea que el 50 % de su catálogo sea contenido original

(procedente de EE. UU) y que al menos se produzcan 80 películas originales y 30 series animadas (género animación) para 2019 [2017 – 2019]. A continuación, comprobaremos si estas metas se cumplen con nuestros datos, para ello se definen a continuación varios KPIs.

8.1.4.1 KPI 4.1: Analizar que el 50% de los contenidos de Netflix sea contenido procedente de EE. UU de 2017 a 2019

Figura 84. Gráfico tipo y cantidad contenidos de EE. UU en Netflix 2017 - 2019 – Elaboración propia

Para ver si se cumplen el KPI 4.1 analizaremos la imagen anterior, podemos ver que entre los años 2017 y 2019 Netflix contiene un total de 229 contenidos, del cual en el gráfico circular podemos ver que aproximadamente un 57% de los contenidos de Netflix son películas y un 43% son series.

Además, podemos ver una tarjeta que nos muestra el porcentaje de contenidos de Netflix cuyo país de producción es EE. UU durante el 2017 y 2019. Como podemos observar el porcentaje de contenidos de EE. UU supera el 50%, por lo que **sí que se cumpliría** el KPI 4.1 que se marcó que el 50% de los contenidos de Netflix sea contenido procedente de EE. UU de 2017 a 2019.

8.1.4.2 KPI 4.2: Analizar que en total en Netflix se producen al menos 80 películas procedente de EE. UU de 2017 a 2019

Figura 85. Gráfico cantidad películas EE. UU de Netflix 2017-2019 – Elaboración propia

En el gráfico anterior, podemos observar la cantidad de películas estadounidenses que Netflix tiene en su catálogo durante los años 2017 y 2019.

Tal y como se aprecia en el gráfico, en los años 2017 y 2018, Netflix lanzó 20 y 19 películas procedentes de EE. UU una cifra muy similar, mientras que en 2019 vemos que las producciones de películas de EE. UU aumentan hasta 28 películas estadounidenses. Esto nos indica que **no se cumpliría** el KPI 4.2 marcado por el informe de Netflix de producir al menos 80 películas estadounidenses en la plataforma Netflix durante este periodo, ya que en total se están produciendo 67 películas estadounidenses en Netflix, una cifra inferior al KPI establecido.

8.1.4.3 KPI 4.3: Analizar que en total en Netflix se producen al menos 30 series animadas de 2017 a 2019.

Figura 86. Gráfico cantidad series animadas de Netflix 2017-2019 – Elaboración propia

En el gráfico anterior se puede ver la cantidad anual de series producidas que pertenecen al género "Animation" las cuales pertenecen al catálogo de Netflix durante los años 2017 y 2019.

Podemos ver que no hay una gran producción de series animadas en Netflix, ya que en 2017 solo hay 3 series animadas, 2018 hubo un aumento en la producción con 10 series animadas y en 2019 hubo una caída con una producción de 7 series animadas. Por lo que **no se cumple** el KPI 4.3 definido por el informe de Netflix, el cual dice de tener al menos 30 series animadas en Netflix ya que en total desde 2017 a 2019 hay 20 series animadas.

8.1.5 Análisis del reconocimiento en premios Oscar

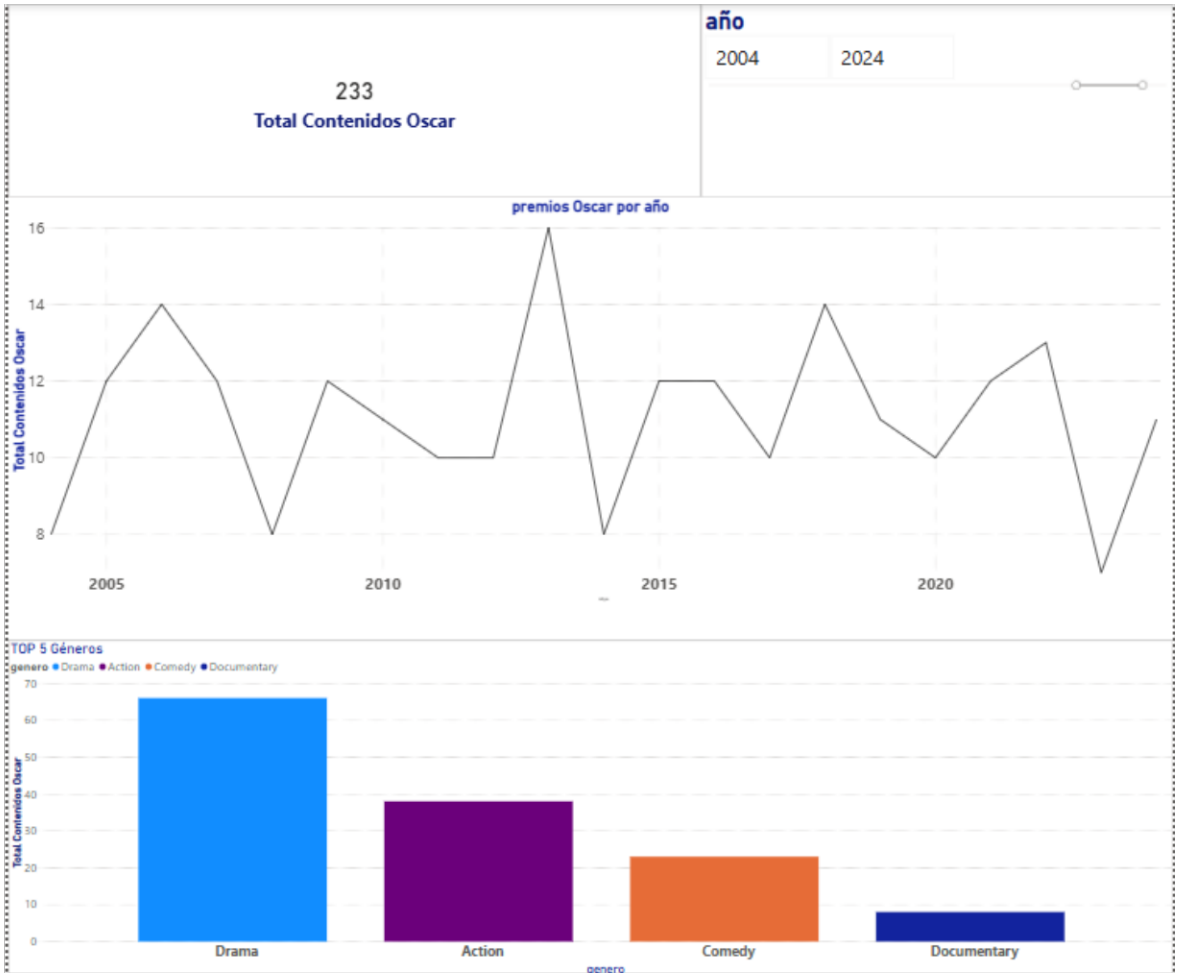


Figura 87. Contenidos ganadores de Oscar 2004-2024 – Elaboración propia.

La imagen anterior muestra la página del dashboard que se ha realizado para analizar los contenidos que han ganado algún premio Oscar. En dicha imagen se incluyó un filtro para que el usuario pueda filtrar la información de los contenidos premiados en un periodo de tiempo concreto.

Además, hay un gráfico de líneas para analizar la evolución temporal en la producción de contenidos premiados con algún Oscar, así como otros aspectos relevantes como los géneros a los que más pertenecen los contenidos premiados.

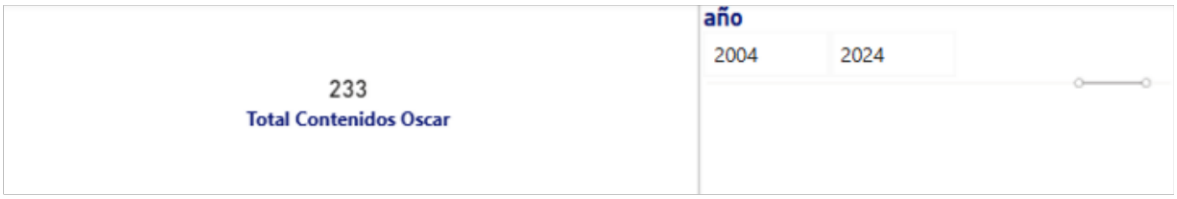


Figura 88. Cantidad contenidos premiados 2004 -2024 – Elaboración propia.

A continuación, se puede ver en la tarjeta la cantidad de contenidos que al menos han ganado algún premio Oscar durante los últimos 20 años [2004 – 2024]. Vemos que hay 233 contenidos premiados.

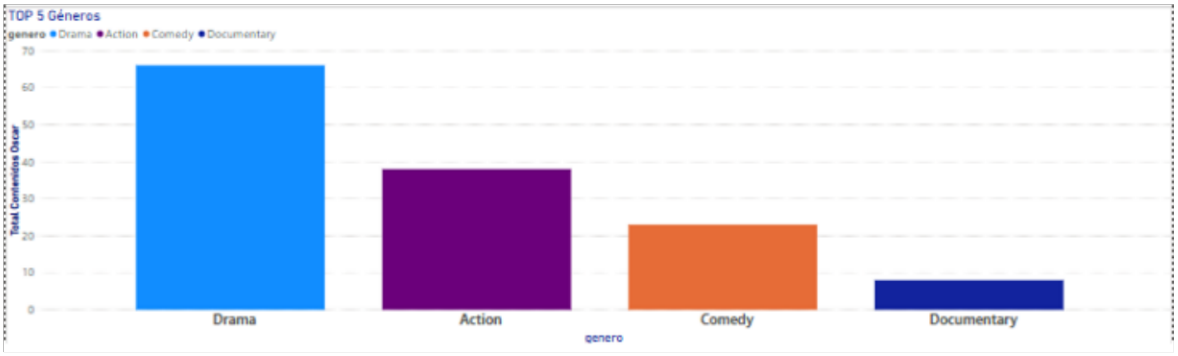


Figura 89. Géneros ganadores de Oscar 2004 – 2024 – Elaboración propia.

En el gráfico anterior, se puede ver los géneros que más se repiten entre los contenidos que han ganado algún premio Oscar durante 2004 y 2024. Podemos observar que el género que destaca ampliamente con más contenidos premiados es "Drama" con 66 contenidos con

algún premio Oscar, lo sigue el género “Action” con 38 contenidos premiados y el tercer género que más se repite es “Comedy” con 23 contenidos.

Se pretende analizar cómo ha evolucionado en los últimos 20 años la producción de los contenidos que han ganado algún Oscar. Como se dijo anteriormente, existen diferentes premios Oscar y es un premio exclusivo para películas. Para ello se definen a continuación dos KPIs.

8.1.5.1 KPI 5.1: Analizar si los contenidos premiados con Oscar aumentan en 5 unidades en los últimos 20 años [2004 – 2024]

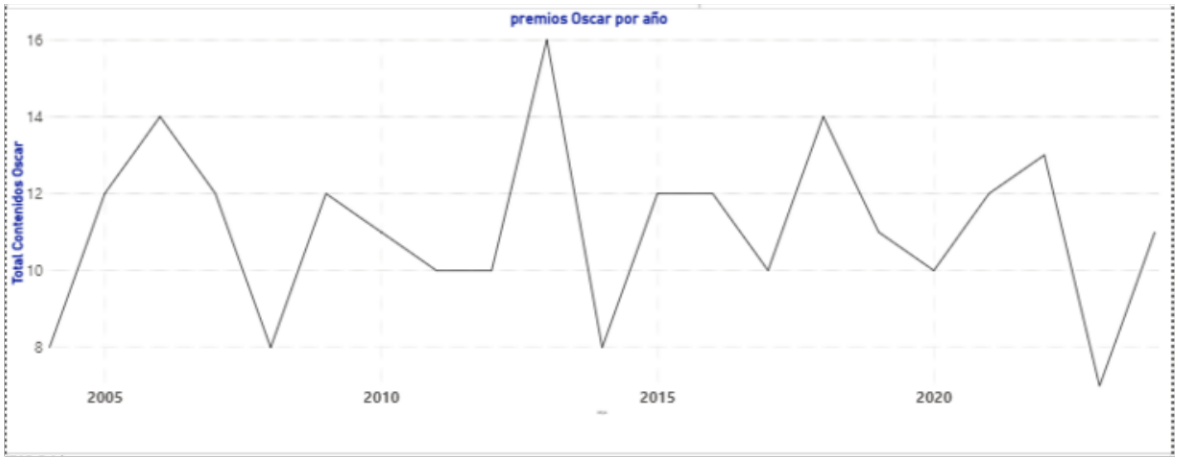


Figura 90. Gráfico evolución producción contenidos con Oscar 2004-2024 – Elaboración propia

Para ver si se cumple el KPI 5.1 analizaremos el gráfico de líneas anterior que muestra la evolución de los contenidos anuales premiados durante 2004 y 2024. Como podemos observar, hay un aumento de los contenidos populares en los últimos 20 años, no obstante, no se llega a incrementar en 5 unidades. Los contenidos premiados con Oscar en 3 unidades, lo que **no cumpliría** el KPI 5.1 establecido que proponía en aumentar al menos los contenidos ganadores de Oscar en 5 unidades.

8.1.5.2 KPI 5.2: Analizar si en total hay al menos 100 contenidos que han ganado algún premio Oscar en los últimos 20 años [2004 – 2024]

Año	Suma de Oscar
2004	8
2005	12
2006	14
2007	12
2008	8
2009	12
2010	11
2011	10
2012	10
2013	16
2014	8
2015	12
2016	12
2017	10
2018	14
2019	11
2020	10
2021	12
2022	13
2023	7
2024	11
Total	**233**

Figura 91. Suma contenidos totales con premios Oscar – Elaboración propia

En la imagen anterior se puede ver la cantidad de contenidos premiados con algún Oscar durante los últimos 20 años. Como podemos observar ha habido un aumento en los contenidos con premios Oscar, siendo 2013 el año con más contenidos premiados con 16 y 2023 el año que menos contenidos premiados hay con 7.

Como podemos observar, **sí que se cumpliría** el KPI 5.2, ya que este establecía que durante el 2004 y 2024 debía haber al menos 100 contenidos premiados, y se cumple ampliamente teniendo un total de 233 contenidos que han recibido algún premio Oscar, como dijimos anteriormente el género que más se repetía entre los contenidos premiados es "Drama".

8.2 Preguntas analíticas

A parte de responder a los objetivos y KPIs definidos en el proyecto, se han planteado una serie de preguntas analíticas, para analizar junto al dashboard y nos ayuden a entender mejor el mercado audiovisual y descubrir patrones.

De esta manera garantizaremos que nuestro proyecto es capaz de llevar a cabo dicho análisis.

Es importante destacar que, en el presupuesto, en la recaudación y en el beneficio, los valores medios no son del todo representativos, ya que este campo se obtuvo por webscraping y en aproximadamente el 90 % de los contenidos este dato se desconoce.

Añadí este dato en el almacén, porque hay algunos contenidos que tengo esa información, y he tomado la decisión de añadirlo en el almacén. En futuros trabajos buscaremos bases de datos alternativas para completar el presupuesto de las películas faltantes.

A continuación, se plantean algunas preguntas analíticas tanto sobre contenidos actuales como otros más antiguos. Para ello nos basaremos en la página principal del dashboard mostrada anteriormente.

8.2.1 ¿Qué géneros dominan en la actualidad?

Anteriormente vimos los géneros que más contenido se han producido en donde destacaban los géneros "Drama", "Documentary" y "Comedy".

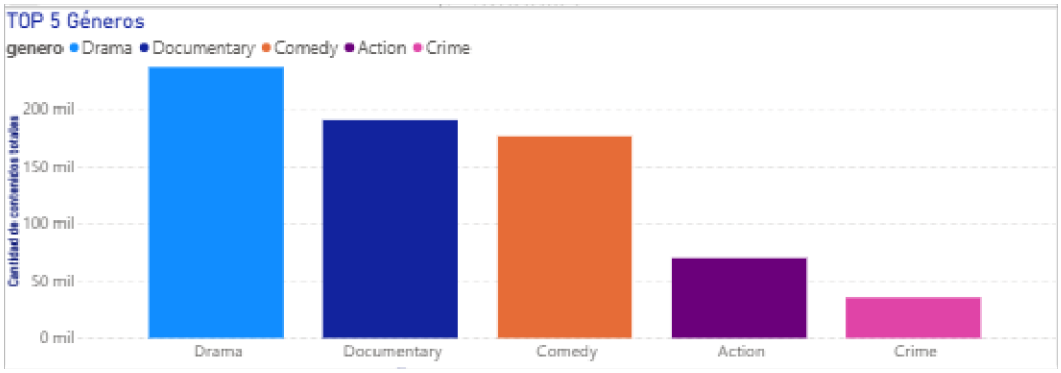


Figura 92. Gráfico géneros más producidos – Elaboración propia

Si analizamos los últimos 10 años (2014 – 2024) podemos observar que los géneros con más contenidos han variado, ya que ahora el género dominante es "Documentary", con aproximadamente un total de 60000 contenidos, mientras que "Drama" que antes era el

género con más contenidos baja al segundo puesto. Además, los contenidos de género "Horror" entran en el top 5 de géneros con más producciones en los últimos 10 años.

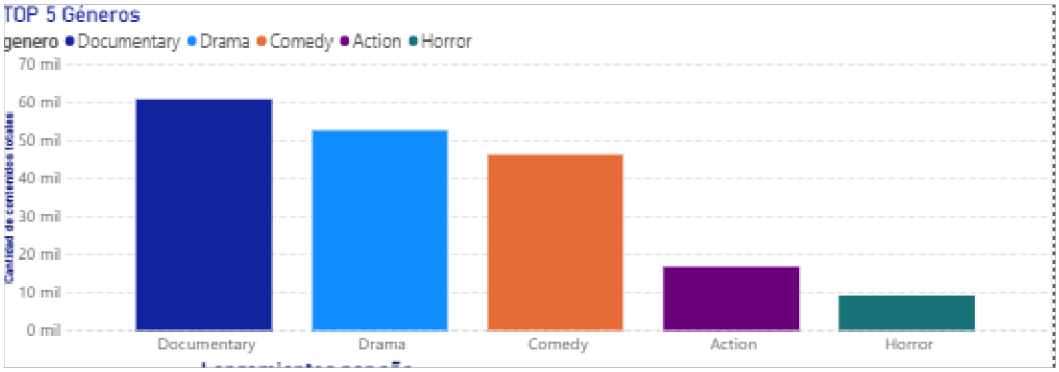


Figura 93. Gráfico géneros populares 2014-2024 – Elaboración propia

8.2.2 ¿Qué géneros tienen una mayor recaudación?

En la siguiente gráfica podemos observar los géneros que más recaudación genera en nuestro almacén de datos.

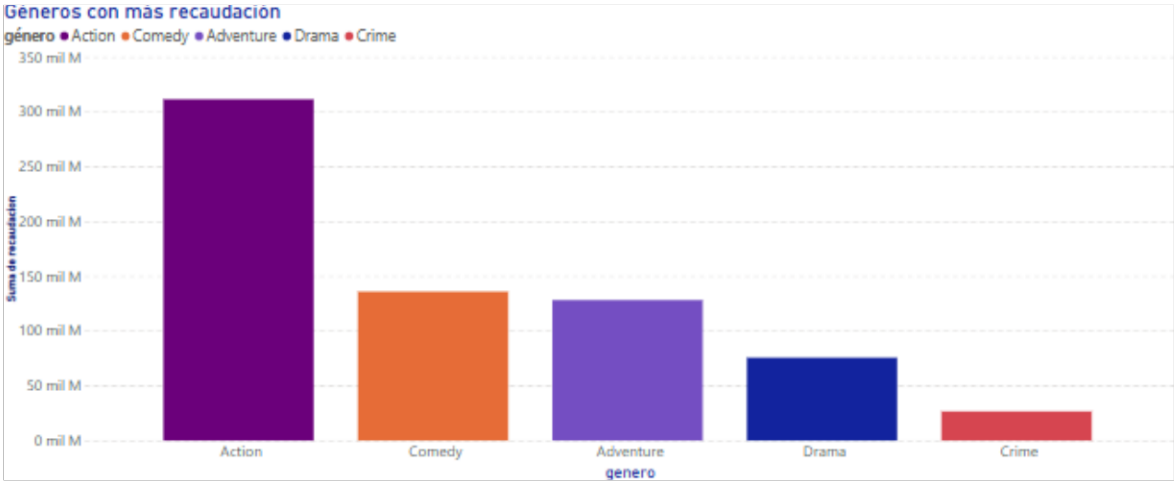


Figura 94. Géneros que producen más ingresos – Elaboración propia

Como podemos observar el género que más ingresos genera es "Action" con una gran diferencia con respecto al resto con más de 300000 millones de euros, el segundo género que más está generando es "Comedy" con aproximadamente 150000 millones de euros y con una cantidad similar en el top 3 encontramos el género "Adventure".

Curiosamente el género que más dinero genera no es el género que más producciones tiene ya que como vimos anteriormente en nuestro almacén el género que más contenidos tiene es

Drama, y es el cuarto género que más dinero genera, mientras que el género Action que es el que más ingresos genera es el cuarto que más producciones tiene.

8.2.3 ¿Qué películas tienen una mayor recaudación y mayor puntuación?

Otra cuestión interesante, es si realmente existe una relación entre el éxito económico de una película (a nivel de recaudación) y su éxito en cuanto al público (a nivel de puntuación). Dicho de otra manera, ¿los contenidos con mayor recaudación son los que más gustan a la audiencia?

Para analizar esto realizaremos 2 visualizaciones, una con el top 10 de contenidos que mayor recaudación tienen, y otras con los contenidos mejor puntuadas por parte de la audiencia.

Es importante destacar que un contenido puede tener una puntuación de 10 y 5 votos, por ejemplo, por lo que para obtener los contenidos con mayor puntuación se han seleccionado aquellos que al menos tengan 1000 votos.

Figura 95. Contenidos con mayor recaudación – Elaboración propia

Como podemos ver el contenido que más recaudación ha tenido es "The Lion King", con casi 2,5 mil millones de euros. Lo sigue muy de cerca "Avatar: The Way of Water," con más de 2 mil millones de euros y en tercer lugar está "Titanic".

En la siguiente gráfica podemos ver los contenidos que mejor puntuación han recibido por parte de la audiencia. En donde podemos ver "The Shawshank Redemption" con una puntuación de 9,3 seguido de "The Godfather" con una puntuación de 9,2, y en tercer lugar vemos a "12 Angry man" con una puntuación de 9.0.

Figura 96. Contenidos más populares en la audiencia – Elaboración propia

Como podemos observar, ninguna de las películas del top 10 que más ha gustado a la audiencia, entra en el top 10 de las películas más taquilleras. Si realizamos una consulta en nuestro almacén de datos de la película que más ha gustado a la audiencia, encontramos que la recaudación es de 27 millones de €, lo que está muy lejos de entrar en el top 10 de las películas más taquilleras, ya que el límite lo marca "Aladin" con 1.5 mil millones de € de recaudación.

En resumen, que una película tenga una buena puntuación por parte del público, no quiere decir que haya sido un "éxito" en taquilla y viceversa.

8.2.4 ¿En qué épocas se realiza más contenidos actualmente?

Anteriormente, en el dashboard histórico observamos que la época que más contenido se realizo es en otoño y en estos 10 últimos años continúa siendo así. Curiosamente este último año 2024 hubo una caída importante de los contenidos en Otoño y el año pasado la época que tiene más contenidos es Primavera con más de 6000 contenidos.

Figura 97. Gráfico épocas lanzamiento 2014-2024 – Elaboración propia

8.2.5 ¿Hay un auge de producciones en nuevos países en los últimos 10 años?

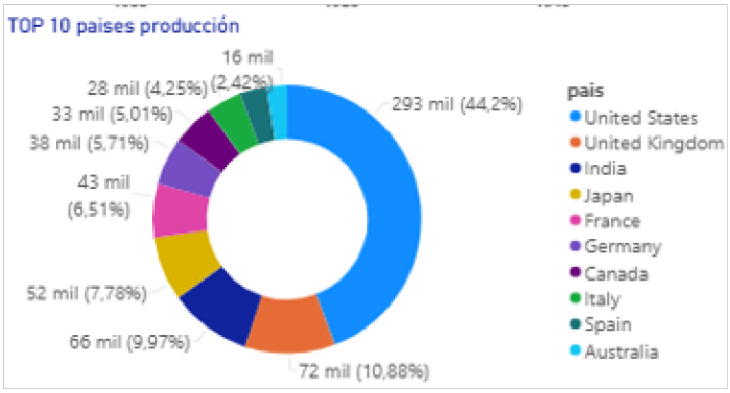


Figura 98. Gráfico países producción populares – Elaboración propia

Como vemos en el gráfico anterior entre los países que más producciones tienen históricamente destaca Estados Unidos con un 44% del contenido total, lo sigue Reino Unido con aproximadamente un 11% del contenido total y en tercer lugar está la India con aproximadamente un 10% del contenido.

En estos últimos 10 años el país que más producciones de contenidos se han realizado continúa siendo EE. UU con casi un 50% de producciones de los contenidos totales, aumentando casi un 6%, por otro lado, India sube al top 2 como segundo país con más producciones con un 11% de contenidos, seguido muy de cerca por Reino Unido que baja hasta el top 3. Cabe destacar que en estos últimos años ha aumentado las producciones de contenido en China, ya que entra en el top 10, superando incluso a países como España e Italia que históricamente estaban en el top 10 de países con más contenidos producidos.

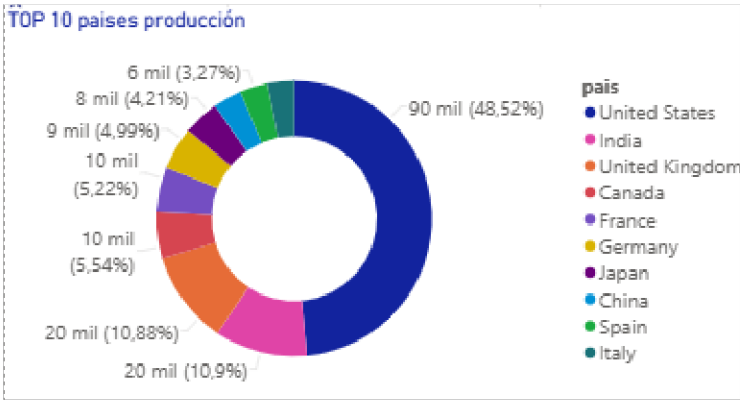


Figura 99. Gráfico países producción populares 2014-2024 – Elaboración propia

8.2.6 ¿De qué países son los actores más populares (estrellas consagradas)?

Con esta pregunta, pretendemos encontrar el origen geográfico de los actores que han alcanzado el estatus de estrellas consagradas en el mundo del contenido audiovisual, de los cuales tenemos registrados en nuestro almacén de datos. Como se dijo anteriormente un actor que tiene la reputación de estrella consagrada es aquel que ha ganado más de 20 premios.

A continuación, podemos ver una gráfica de los diferentes países de procedencia de dichos actores, junto al número total de actores consagrados de ese país y su porcentaje.

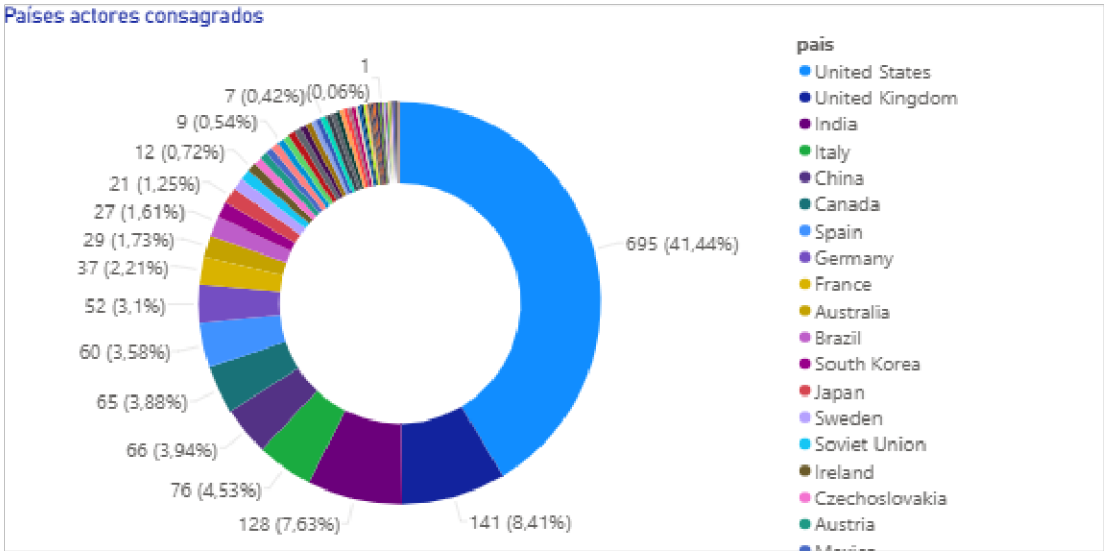


Figura 100. Países actores consagrados – Elaboración propia

Como podemos observar casi la mitad de los actores consagrados proceden de EE. UU. Reino Unido ocupa el segundo puesto a bastante distancia, sus actores consagrados representan algo más de un 8%. En tercer lugar, encontramos a la India con algo menos de un 8%. Luego lo siguen otros muchos países con menor peso como Italia, China, Canadá o España.

8.2.7 ¿A qué grupo de edad pertenecen los actores conocidos?

A continuación, analizaremos los grupos de edad, a los que pertenecen los actores que son considerados "conocidos". Para ello es importante recordar que en el punto 6 en la dimensión de actor, ya realizamos un histograma, para categorizar la edad en diferentes grupos. Los grupos de edad son:

Menor de edad: 0 – 17 años

Joven: 19 – 30 años

Adulto: 31 – 60 años

Adulto mayor: +60 años

En la siguiente gráfica, podemos observar de los últimos 10 años (2014 – 2024) a que grupo de edad pertenecen los 10 actores que aparecen en las producciones mejor puntuadas por parte de la audiencia.

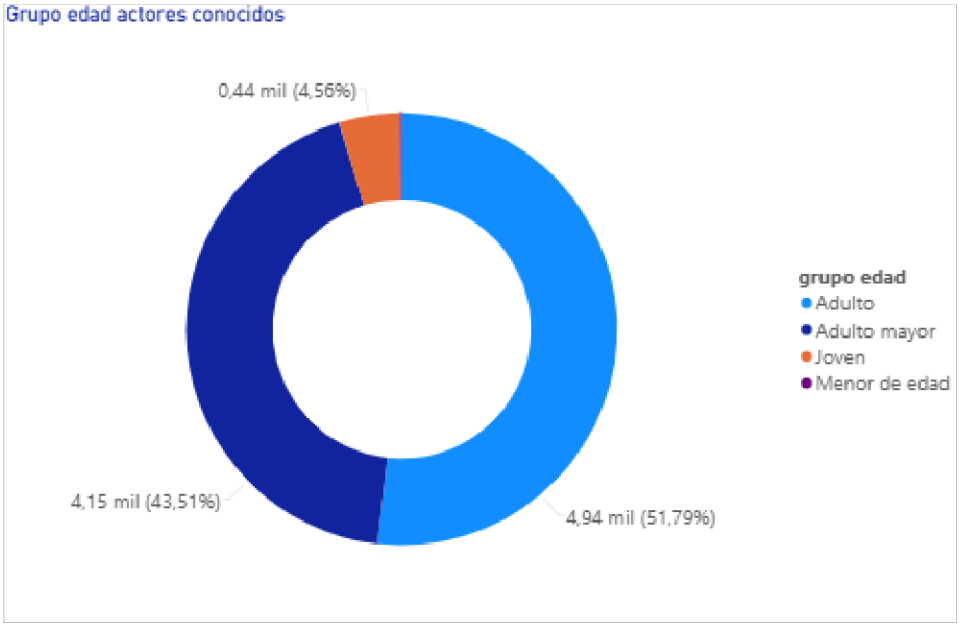


Figura 101. Grupo edad actores conocidos entre 2014 y 2024 – Elaboración propia

Como podemos observar, aproximadamente poco más de la mitad de los actores conocidos, pertenecen al grupo de edad "Adulto", es decir que tienen una edad entre 31 y 60 años, mientras que el 44% son "Adultos mayores", es decir tienen más de 60 años, por otro lado, un 4% son actores "Jóvenes", tienen entre 19 y 30 años y solamente un 0,15% son menores de edad.

8.2.8 ¿Un director ganador de algún Oscar influye en la puntuación de la película?

Con esta pregunta, se busca analizar si la participación de un director galardonado con algún premio Oscar tiene un impacto significativo en la valoración media del contenido por parte de la audiencia. Para ello, se ha comparado la puntuación media de los contenidos que son dirigidos por directores que han ganado algún premio Oscar, que está representado en la barra de la izquierda de color azul claro, frente a los dirigidos por aquellos que no han recibido ningún premio Oscar, representado en la barra de la derecha en azul oscuro.

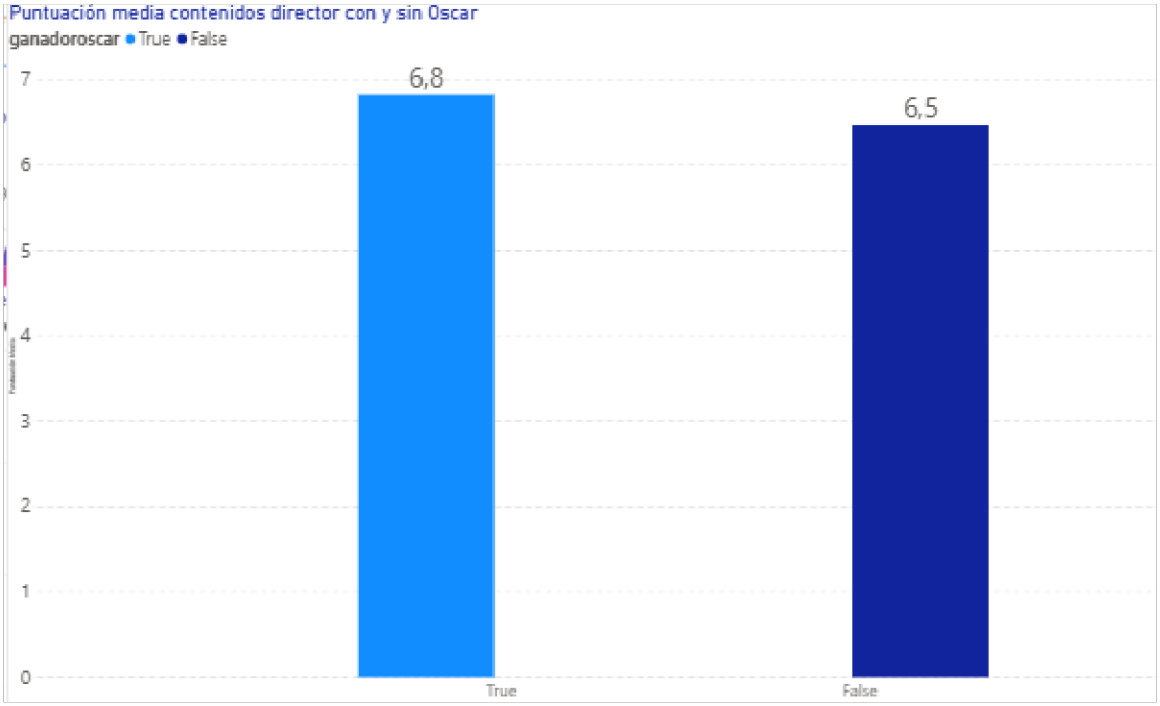


Figura 102. Puntuación media contenidos cuando un director tiene y no tiene Oscar – Elaboración propia

Como podemos ver en el gráfico, los contenidos que son dirigidos por un director ganador de algún Oscar tienen una puntuación media de 6.8 mientras que los que son dirigidos por directores sin Oscar tienen puntuación media ligeramente inferior de 6,5.

Esta diferencia, no es especialmente significativa, lo que nos lleva a pensar que contar con en un contenido con un actor ganador de algún Oscar si puede influir positivamente en la valoración de un contenido, pero no de una forma muy notable.

Es probable que otros factores, como el reparto, el guion, o el género tengan un peso mayor en la valoración del público.

8.2.9 ¿Qué idiomas originales son los más frecuentes en los contenidos?

Con esta pregunta buscamos conocer los idiomas originales más frecuentes en todos los contenidos del almacén de datos.

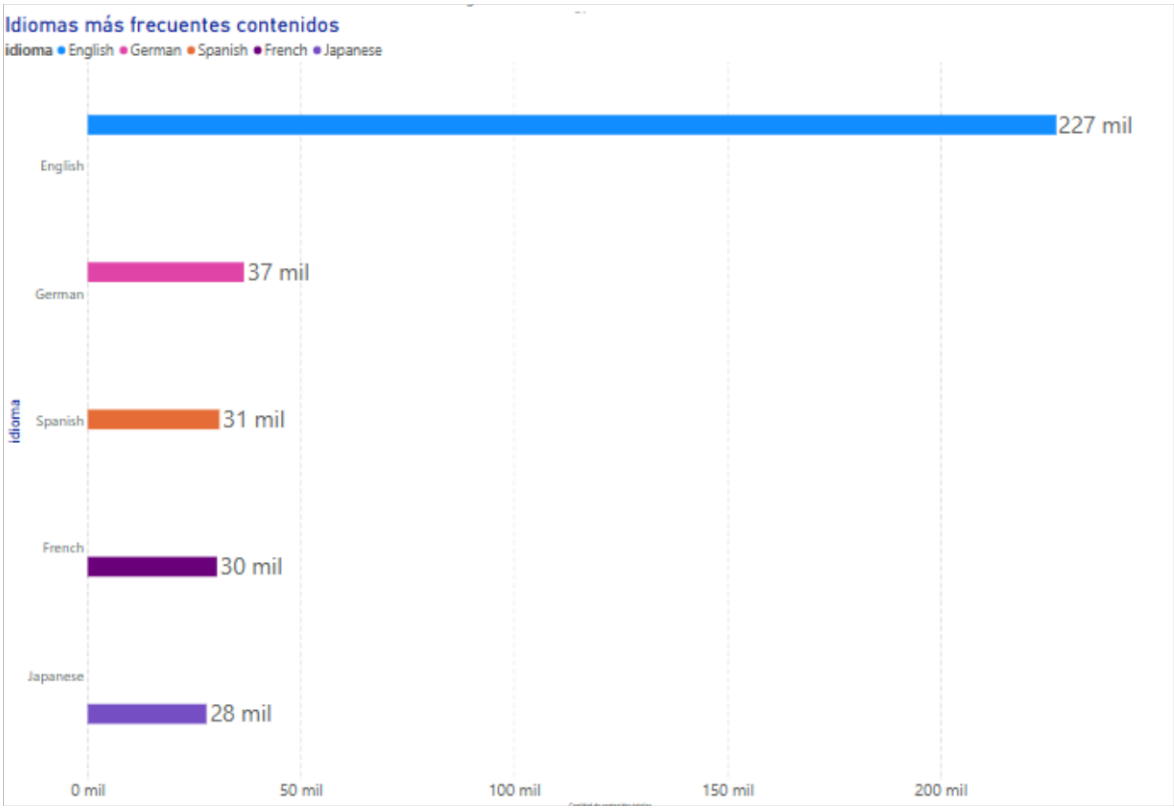


Figura 103. Idiomas originales más frecuentes contenidos – Elaboración propia

Como podemos observar el inglés es el idioma más dominante con mucha diferencia en cuanto a lengua original de un contenido con 227000 contenidos aproximadamente. El segundo idioma original que más aparece en los contenidos es el alemán con 37000 contenidos. Luego vemos otros idiomas con cifras de contenido muy parecidas, como el español, el francés y el japonés.

8.2.10 ¿Qué plataformas tienen los contenidos mejores valorados?

Para responder a esta pregunta se han analizado la puntuación media de todos los contenidos en cada plataforma, considerando únicamente las tres plataformas claramente identificadas Netflix, AppleTV+ y PrimeVideo y excluyendo tanto las entradas sin plataforma asignada como las que pertenecen a "Multiplataforma".

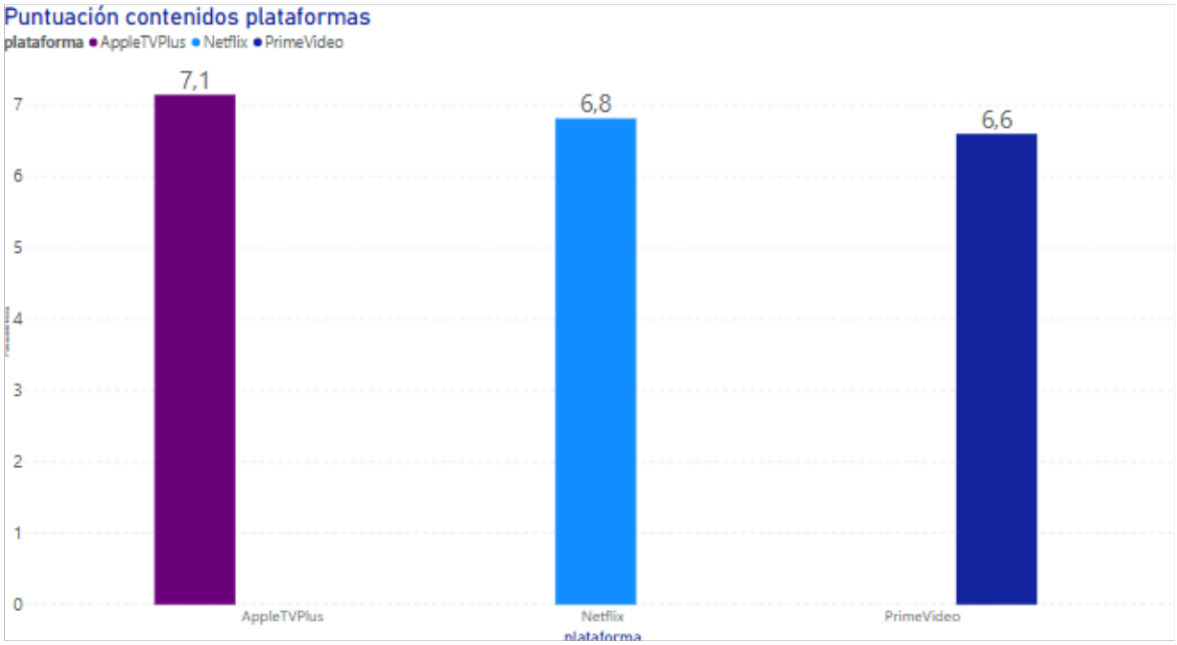


Figura 104. Gráfica puntuación por plataforma – Elaboración propia

En el gráfico de barras resultado podemos observar que:

AppleTV+ lidera la tabla con una puntuación media aproximada de 7,1. Netflix se sitúa en lo sigue, con una media de 6,8 puntos y PrimeVideo queda en tercer lugar, con una puntuación media de 6,6 puntos.

Estos resultados indican que, los contenidos de AppleTV+ tienen la valoración ligeramente más alta por parte de la audiencia. Netflix también muestra una buena puntuación, aunque ligeramente por debajo de AppleTV+. Finalmente, PrimeVideo, con un promedio de 6,6, sigue de cerca a Netflix.

La diferencia entre AppleTV+ con una valoración de 7,1 y PrimeVideo valorado con 6,6. Esto indica que, aunque todas las plataformas ofrecen contenidos bien puntuados. AppleTV+ destaca ligeramente por el resto de las plataformas en cuanto a puntuación por parte de la audiencia.

8.2.11 ¿Cómo ha crecido el volumen de contenidos de las plataformas de streaming en los últimos 20 años?

Para responder a esta pregunta analizaremos la cantidad de contenidos que ha tenido cada plataforma de streaming desde 2004 hasta 2024. Esta evolución la podemos observar en la

siguiente gráfica, en la que se representan los catálogos de Netflix, Prime Video y AppleTV+ en los últimos 20 años.

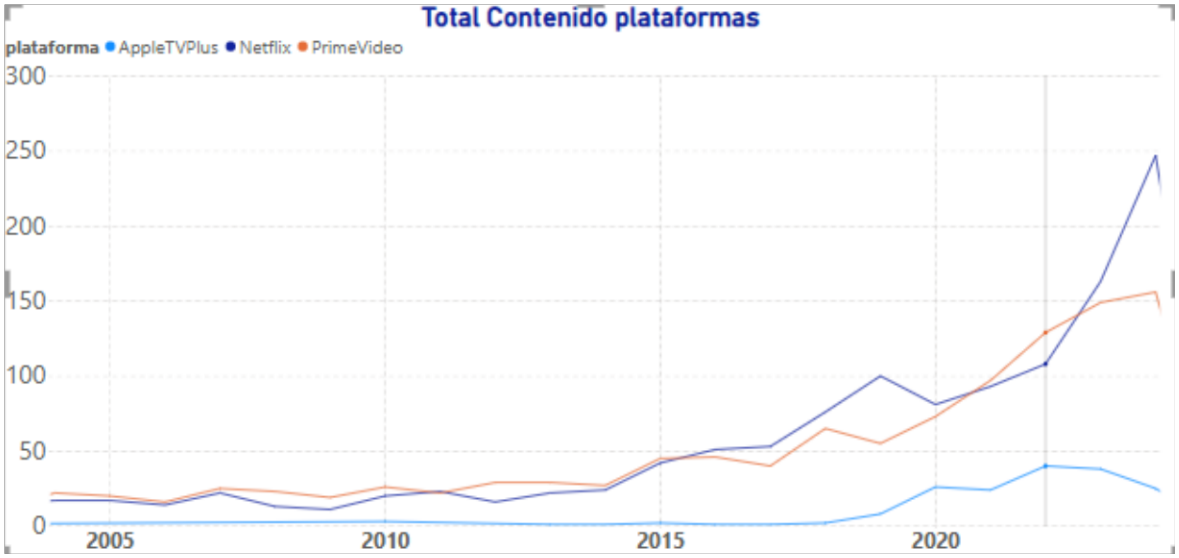


Figura 105. Gráfica cantidad de contenidos por plataforma – Elaboración propia

En la gráfica podemos ver cómo desde el 2004 hasta el 2014, el crecimiento de contenidos de las plataformas fue bastante moderado. No obstante, a partir de 2015, el crecimiento comienza a acelerarse, sobre todo en Netflix, que duplica su catálogo entre 2016 y 2019. Prime Video tiene una evolución similar, aunque algo más lenta.

Por otro lado, AppleTV+ se empieza a notar su crecimiento de contenidos en 2019 ya que años anteriores apenas tenía contenidos.

A partir del 2020, coincidiendo con la pandemia podemos ver que hay un aumento muy notorio en las plataformas Netflix y Prime Video. Netflix se consolida como la plataforma con más contenidos con aproximadamente 250 contenidos mientras que Prime Video tiene 150.

Podemos observar cómo el año de la pandemia supuso un punto de inflexión en cuanto al consumo de contenidos en plataformas de streaming.

9. Conclusiones

La principal problemática es que a día de hoy no hay páginas que permitan realizar análisis detallados con información masiva de contenidos audiovisuales, es por ello que en este proyecto se ha abordado el reto de analizar a gran escala diferentes datos de millones de contenidos audiovisuales.

Las fuentes de información estaban dispersas, los datos se obtuvieron de diferentes páginas y en muchos casos, mucha información está restringida (IMDb, JustWatch), lo que obliga a utilizar técnicas adicionales para obtener esa información.

A partir de la problemática se proponen varios objetivos con sus KPIs correspondientes para analizar el contexto de la producción de diferentes contenidos.

En cuanto al objetivo de analizar las producciones nacionales, se plantearon dos KPIs basados en el Plan de Fomento del Cine Español 2021–2024: comprobar un crecimiento del +30 % tanto en películas como en series españolas entre 2021 y 2024. Los resultados muestran que, lejos de cumplirse, las películas nacionales han experimentado una disminución de un −30 % y las series un −62 % en ese periodo, situándose muy por debajo de los umbrales marcados en el KPI 1.1 y KPI 1.2.

Cuando se analizó los contenidos populares se pudo observar una evolución positiva en la producción de películas, con un incremento acumulado del +173 % en los últimos 20 años, lo que supera de largo el límite máximo de caída permitido (KPI 2.1) que marcaba que los contenidos populares no cayeran más de un -15%. Además, más del 10 % de las series populares corresponden al género Drama, validando con creces el KPI 2.2.

Al analizar los lanzamientos en fechas clave en España y en EE. UU, se definieron cuatro KPIs: aumento mínimo del +15 % en estrenos de Navidad que corresponde al KPI 3.1 y +20 % en Verano para España que corresponde al KPI 3.2, y del +20 % en Otoño y Primavera para EE. UU, KPI 3.3 y KPI 3.4 respectivamente. Todos ellos se cumplieron de sobra, con incrementos acumulados superiores al doble de los umbrales definidos.

Para el análisis de Netflix, se marcaron los KPIs definidos por un informe de Netflix el cual fijó la meta de que el 50 % de su catálogo fuese contenido estadounidense y la producción de al menos 80 películas y 30 series animadas entre 2017 y 2019. Si bien se alcanzó aproximadamente un 50 % de títulos de EE. UU que es el KPI 4.1, el total de películas es de 67 y el total series animadas es de 20 por lo que no se cumplen los KPIs 4.2 y 4.3.

En cuanto al reconocimiento en premios Oscar se definieron los siguientes KPIs: incremento mínimo de +5 producciones ganadoras que es el KPI 5.1, y al menos 100 contenidos con algún Oscar que es el KPI 5.2. Se superó el umbral de 100 contenidos, pero no se alcanzó un aumento de +5 unidades en el total de ganadores, y el conteo de categorías distintas quedó por debajo de las cinco establecidas.

Como se puede observar se cumplieron la mayoría de KPIs, no obstante, hay algunos como el KPI 1.1 y 1.2 definidos basándose en el Plan de Fomento del Cine Español que quedaron muy lejos del umbral marcado incluso llegando a haber una disminución de contenidos tanto en películas como en series.

Se diseñó una arquitectura Big Data capaz de integrar información masiva procedente de diferentes fuentes como IMDb, para ello se descargaron los datasets oficiales de la página de "Non Commercial Datasets" y posteriormente se utilizaron técnicas de webscraping para ampliar la información de los contenidos. Por otro lado, para obtener la información de JustWatch se obtuvo totalmente mediante webscraping y usando librerías como BeautifulSoup que permite acceder al contenido HTML de una web y Selenium WebDriver que permite simular acciones de usuario en una web.

Para procesar la información se ha utilizado Apache Spark con el lenguaje de programación Scala debido a su eficiencia y procesado de datos en memoria frente a otras herramientas. Por otra parte, para almacenar datos se utilizó Hive que usa el lenguaje SQL debido a su escalabilidad con grandes volúmenes de datos. Mientras que para visualizar los datos se ha utilizado Power BI ya que permite crear dashboards con varias páginas y en dichas páginas incluir filtros para que el usuario pueda ajustar los datos según sus necesidades.

Podemos observar que, tras el análisis realizado las producciones nacionales se han reducido en los últimos años y que, por el contrario, las plataformas de streaming como Netflix o Prime Video han aumentado exponencialmente su catálogo de contenidos.

Por otro lado, la solución implementada además de permitir analizar diferentes aspectos de los contenidos audiovisuales, así como su evolución en el tiempo, garantiza escalabilidad ya que en un futuro se podrá ampliar las fuentes de datos y añadir nuevos nodos para acelerar la ejecución de los procesos y evitar cuellos de botellas.

10. Posibles mejoras futuras

En el proyecto realizado hay algunos aspectos a mejorar que podrían enriquecer y hacer más preciso y eficiente los análisis realizados.

10.1 Enriquecer la información con fuentes externas

Actualmente, casi el 90% de los contenidos se desconoce el valor de las medidas presupuesto y la recaudación. Por lo que los análisis que realizamos con estos campos no son del todo representativos.

Se podría integrar información adicional con otros datasets públicos o APIs de otras páginas que contengan información sobre contenidos audiovisuales y a la vez hacer que los análisis realizados sobre esos campos sean más fiables.

Lo mismo se puede hacer con otros campos que se desconoce el valor como fechas de nacimiento de actores, países de producción, etc.

10.2 Escalabilidad con un clúster distribuido

Ahora mismo, todo se está ejecutando sobre un mismo servidor, esto puede acabar produciendo un cuello de botella si estamos trabajando con una cantidad masiva de datos. Si creamos un clúster Hadoop con varias máquinas, se podría repartir la carga de trabajo en varios nodos. De esta manera, en vez de ejecutar todo en una misma máquina, cada nodo se encargaría de una parte, así se reduciría de forma notable los tiempos de carga y transformación de datos.

10.3 Inteligencia artificial

El uso de la inteligencia artificial podría ayudarnos a anticipar cómo va a ser la recepción de un contenido antes de su estreno. Se puede obtener información de diferentes redes sociales como opiniones de Twitter, comentarios y valoraciones en Reddit para analizar la valoración del público en tiempo real. Con esos datos, usando la IA se podría predecir si un estreno de un contenido va a gustar más o menos.

11. Opinión personal

A nivel personal, este proyecto me ha servido como un ejercicio de disciplina y de aprendizaje. El hecho de dividir el proyecto en diferentes fases, como el diseño del almacén de datos, el desarrollo de los ETLs, fue vital para poder ir avanzando de manera ordenada e ir evitando posibles bloqueos en el proyecto. He podido comprobar, que en este tipo de proyectos BI, con gran envergadura, el proceso de planificación e ir iterando las tareas juega un papel fundamental a la hora de poder ir avanzando en el proyecto de manera eficiente.

Por otro lado, este proyecto me ha ayudado a entender más lo importante que es la analítica, no solo como recurso a la hora de poder tomar decisiones en una empresa que quiera sacar algún contenido, sino también para entender mejor los patrones de consumo cultural y poderse anticipar a las tendencias. Aprender a usar herramientas como Spark y Hive, además de aprender la técnica de webscraping junto a librerías como BeautifulSoup y Selenium WebDriver para recoger cierta información de páginas webs, para trabajar con millones de datos y posteriormente poder visualizarlos en una herramienta que permite crear dashboards como Power BI, refuerza más mi idea de que el Business Intelligence es una parte fundamental en cualquier empresa.

Para concluir, estoy satisfecho con haber podido sintetizar en una abrumadora cantidad de información de contenidos audiovisuales en diferentes fuentes, desde presupuesto, estrenos, información sobre personas relacionadas con la industria audiovisual, etc. Estos conocimientos aprendidos, suponen un mejor punto de partida para poder afrontar posibles futuras implementaciones en el futuro, como por ejemplo explorar nuevas fuentes para completar aquellos datos del almacén que estén muy incompletos como el presupuesto, incorporar IA para poder intuir futuros éxitos de contenidos.

Referencias

[1] Repositorio Trello personal https://trello.com/b/QPfpEPve/tfg-peliculas

[2] ICAA Informe 2024. https://spainaudiovisualhub.mineco.gob.es/content/dam/seteleco-hub-audiovisual/resources/pdf/informe_2024/2024_2_Informe_Spain_Audiovisual.pdf

[3] Informe Netflix Objetivos 2017 (Accessed: 2025-06-2025). https://en.wikipedia.org/wiki/Netflix%2C_Inc.?utm_

[4] Motion Picture Association (MPA). (2021). Theatrical and Home Entertainment Market Environment Report.

[5] Statista. (2025). *Ranking mundial de los países con mayor número de salas de cine*. https://es.statista.com/estadisticas/1011196/ranking-mundial-de-los-paises-con-mayor-numero-de-salas-de-cine/ (Accessed: 2025-02-06)

[6] Power, D. J. (2007). *A Brief History of Decision Support Systems*. DSSResources.COM. http://dssresources.com/history/dsshistory.html (Accessed: 2025-02-07)

[7] Tableau. (s.f.). *¿Qué es Business Intelligence?* https://www.tableau.com/es-es/learn/articles/business-intelligence (Accessed: 2025-02-07)

[8] Kimball, R., & Caserta, J. (2011). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley.

[9] ESAN University. (2022). *Business intelligence: 6 beneficios de la inteligencia empresarial*. Recuperado de https://www.esan.edu.pe/conexion-esan/business-intelligence-6-beneficios-de-la-inteligencia-empresarial (Accessed: 2025-02-08)

[10] Kimball, R., & Ross, M. (2011). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). Wiley.

[11] Qlik. (n.d.). *ETL vs ELT* https://www.qlik.com/us/etl/etl-vs-elt (Accessed: 2025-02-09)

[12] Inmon, W. H. (1996). *Building the Data Warehouse*. Wiley.

[13] Kimball, R. (1998). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley.

[14] Amazon Web Services. (s.f.). What is a Data Lake? https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/ (Accessed: 2025-02-11)

[15] Alation. (2023). *What is a Data Swamp? Data Swamp vs. Data Lake vs. Data Lakehouse.* https://www.alation.com/blog/data-swamp-lake-lakehouse/ (Accessed: 2025-02-12)

[16] IBM ¿Que es OLAP? https://www.ibm.com/es-es/topics/olap (Accesed: 2025-02-13)

[17] The V's of Big Data https://www.ibm.com/think/topics/big-data (Accesed: 2025-02-21)

[18] Matt Turck (2023). The 2023 MAD (Machine Learning, AI, and Data) Landscape https://mattturck.com/mad2023/

[19] IBM - Apache Spark (Accessed 2024-07-17) https://www.ibm.com/es-es/topics/apache-spark

[20] Apache Spark - MLlib: Machine Learning Library (Accessed 2024-07-17) https://spark.apache.org/mllib/

[21] Apache Spark - Spark Streaming Programming Guide (Accessed 2024-07-18) https://spark.apache.org/docs/latest/streaming-programming-guide.html

[22] Apache Spark - Spark SQL, DataFrames and Datasets Guide (Accessed 2024-07-20) https://spark.apache.org/docs/latest/sql-programming-guide.html

[23] Apache Spark - GraphX Programming Guide (Accessed 2024-07-20) https://spark.apache.org/docs/latest/graphx-programming-guide.html

[24] ¿What is databricks? (Accessed 2025-05-23) https://learn.microsoft.com/es-es/azure/databricks/introduction/

[25] Apache Flink (Accessed 2025-03-02) https://flink.apache.org/

[26] Apache Beam (Accessed 2025-03-02) https://beam.apache.org/

[27] Apache Iceberg (Accessed 2025-03-02) https://iceberg.apache.org/

[28] IBM - Apache Kafka Overview (Accessed 2024-07-26) https://www.ibm.com/es-es/topics/apache-kafka

[29] IBM - "Apache Hive" (Accessed 2024-08-07) https://www.ibm.com/analytics/hive

[30] IBM - Apache Cassandra Monitoring: Overview (Accessed 2024-07-29) https://www.ibm.com/es-es/products/instana/supported-technologies/cassandra-monitoring

[31] IBM - ¿Qué es MongoDB?" (Accessed 2024-07-31) https://www.ibm.com/es-es/topics/mongodb.

[32] IBM - "¿Qué es HBase?" (Accessed 2024-08-02) https://www.ibm.com/es-es/topics/hbase.

[33] Cassandra vs MongoDB vs HBase (Accessed 2024-08-02) https://www.aimoretechnologies.com/cassandra-vs-mongodb-vs-hbase

[34] IBM. Hadoop: Un Marco para el Procesamiento de Datos en Gran Escala. (Accessed 2024-07-13) https://www.ibm.com/es-es/topics/hadoop

[35] IBM - HDFS (Accessed 2024-07-15) https://www.ibm.com/es-es/topics/hdfs

[36] IBM - MapReduce (Accessed 2024-07-15) https://www.ibm.com/topics/mapreduce

[37] IBM - YARN (Yet Another Resource Negotiator) (Accessed 2024-07-16) https://www.ibm.com/es-es/products/instana/supported-technologies/hadoop-yarn-monitoring

[38] Apache Ambari. Apache Software Foundation. (Accessed 2025-04-23) https://ambari.apache.org/

[39] Pentaho. (s.f.). Pentaho – Data Integration & Analytics. (Accessed 2025-04-23)

[40] IBM. (s.f.). StreamSets DataOps Platform. (Accessed 2025-04-23) https://www.ibm.com/products/streamsets?utm_content=SRCWW&p1=Search&p4=43700081676146443&p5=p&p9=5870000882779881 4&gad_source=1&gbraid=0AAAAA-oJ0vJN8lozow9DhJwvyznP6diey&gclid=CjwKCAjwn6LABhBSEiwAsNJrjuaDjah1ixpQ0fmPXR1RQFUSjdQHNpnbllptlWIW_E1eyiU6uajIzRoCgqgQAvD_BwE&gclsrc=aw.ds

[41] IBM. SQL Server Management Studio. (Accessed 2024-09-16) https://learn.microsoft.com/es-es/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver16

[42] Apache Kylin. (s.f.). Apache Kylin: Real-Time OLAP on Hadoop. (Accessed 2025-03-02) https://kylin.apache.org/

[43] Apache Zeppelin: (Accessed 2024-09-02) https://www.dremio.com/wiki/apache-zeppelin

[44] IBM Knowledge Accelerators - Jupyter Notebook. (Accessed 2024-09-03) https://www.ibm.com/docs/en/knowledge-accelerators/1.0.0?topic=catalog-jupyter-notebook

[45] Databricks Notebooks. (s.f.). Databricks Unified Data Analytics Platform. (Accessed 2025-03-04) https://databricks.com/

[46] Microsoft Visual Studio. (s.f.). Visual Studio – IDE for Windows and Mac. (Accessed 2025-03-04) https://visualstudio.microsoft.com/

[47] Oracle. VirtualBox. (Accessed 2025-04-23) https://www.virtualbox.org/

[48] IBM Docker." (Accessed 2024-09-04) https://www.ibm.com/es-es/topics/docker

[49] IBM. Power BI. (Accessed 2024-09-16) https://www.ibm.com/docs/es/spfc?topic=service-power-bi

[50] Tableau Software. (s.f.). Tableau – Business Intelligence & Data Visualization. (Accessed 2024-09-04) https://www.tableau.com/

[51] Qlik Sense. (s.f.). Qlik Sense – Data Analytics and Business Intelligence. (Accessed 2024-09-04) https://www.qlik.com/us/products/qlik-sense

[52] IMDb. (2025). IMDb Non-Commercial Datasets. (Accessed 2025-04-08) https://developer.imdb.com/non-commercial-datasets/

[53] Diseño físico Hive – Repositorio Github personal https://github.com/AlexSeguii/TFG/blob/main/diseños/diseño_físico_hive.sql

[54] Repositorio GitHub personal https://github.com/AlexSeguii/TFG/tree/main

[55] Ficheros webscraping – Reposisorio GitHub personal https://github.com/AlexSeguii/TFG/tree/main/webscraping

Anexo. Instalación del software

Para instalar y ejecutar las diferentes herramientas, utilizaré el sistema operativo Ubuntu. He seleccionado este sistema operativo porque funciona bien como máquina virtual con pocos recursos, además de que estoy familiarizado con él.

1. Apache Spark

Para la realizar la instalación de Spark, en primer lugar, deberemos instalar Java. Para ello ejecutamos el siguiente comando

> apt install openjdk-8-jdk

Posteriormente comprobamos que se esté instalado java.

```
asa117@alejandro:~$ java -version
openjdk version "1.8.0_422"
OpenJDK Runtime Environment (build 1.8.0_422-8u422-b05-1~24.04-b05)
OpenJDK 64-Bit Server VM (build 25.422-b05, mixed mode)
```

Figura Anexo 1. Comprobación versión java

Posteriormente procedemos a la instalación de Spark. Para ello accedemos a la página oficial y descargamos el paquete con el siguiente comando:

```
asa117@alejandro:/opt$ sudo wget https://dlcdn.apache.org/spark/spark-3.5.3/spar
k-3.5.3-bin-hadoop3.tgz
[sudo] contraseña para asa117:
--2024-09-27 10:12:41--  https://dlcdn.apache.org/spark/spark-3.5.3/spark-3.5.3-
bin-hadoop3.tgz
Resolviendo dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Conectando con dlcdn.apache.org (dlcdn.apache.org)[151.101.2.132]:443... conecta
do.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 400864419 (382M) [application/x-gzip]
Guardando como: 'spark-3.5.3-bin-hadoop3.tgz'
```

Figura Anexo 2. Descarga Spark

Una vez instalado en la carpeta opt lo descomprimimos con el siguiente comando:

```
asa117@alejandro:/opt$ sudo tar xzvf spark-3.5.3-bin-hadoop3.tgz
spark-3.5.3-bin-hadoop3/
spark-3.5.3-bin-hadoop3/data/
spark-3.5.3-bin-hadoop3/data/graphx/
spark-3.5.3-bin-hadoop3/data/graphx/users.txt
spark-3.5.3-bin-hadoop3/data/graphx/followers.txt
spark-3.5.3-bin-hadoop3/data/mllib/
```

Figura Anexo 3. Descomprimir paquete Spark

A continuación, tendremos que añadir las variables de entorno en el fichero .bashrc en la carpeta root. También ejecutaremos el comando source .bashrc para que se guarden los cambios.

```
root@alejandro: ~
  GNU nano 7.2                    .bashrc *
# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
#if [ -f /etc/bash_completion ] && ! shopt -oq posix; then
#    . /etc/bash_completion
#fi

export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
export PYSPARK_PYTHON=/usr/bin/python3
```

Figura Anexo 4. Configuración de variables de entorno

Después de instalar las variables de entorno ya podemos ejecutar Spark y comprobar que funcione, ya sea por Python o por Scala. Cabe destacar que el puerto que usa Spark es el 8080.

```
root@alejandro:~# pyspark
Python 3.12.3 (main, Sep 11 2024, 14:17:37) [GCC 13.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
24/09/27 10:48:32 WARN Utils: Your hostname, alejandro resolves to a loopback addr
ess: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
24/09/27 10:48:32 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another ad
dress
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(
newLevel).
24/09/27 10:48:33 WARN NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.3
      /_/

Using Python version 3.12.3 (main, Sep 11 2024 14:17:37)
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1727426914504).
SparkSession available as 'spark'.
```

Figura Anexo 5. Ejecución comando pyspark

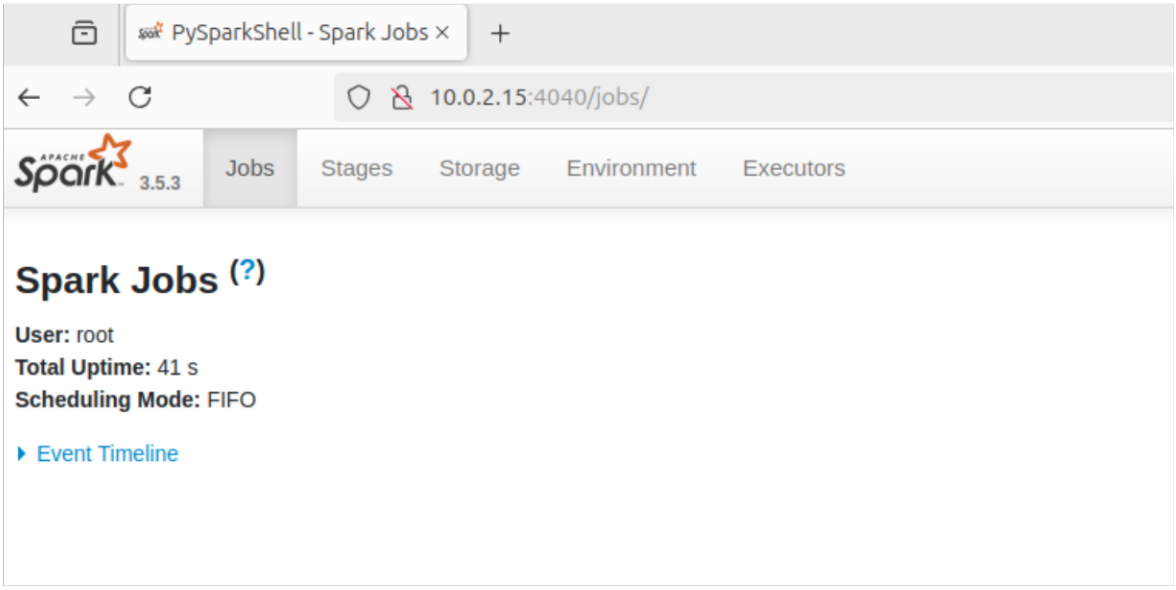


Figura Anexo 6. Funcionamiento de Spark

2. Apache Hadoop

Para instalar Hadoop entramos en la página oficial de Hadoop y seguimos las instrucciones puestas.

En primer lugar, modificamos el fichero .bashrc y añadimos las siguientes líneas:

```
asa117@alejandro: ~
GNU nano 7.2                    .bashrc
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:/usr/lib/jvm/java-8-openjdk-amd64/bin
export HADOOP_HOME=~/hadoop-3.2.3/
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_STREAMING=$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.>
export HADOOP_LOG_DIR=$HADOOP_HOME/logs
export PDSH_RCMD_TYPE=ssh
```

Figura Anexo 7. Fichero .bashrc

A continuación, se procederá a instalar el paquete SSH (Secure Shell). Este comando instala el cliente y el servidor SSH, permitiendo tanto la conexión a otros servidores como la recepción de conexiones remotas en el mismo equipo.

```
asa117@alejandro:~$ sudo apt-get install ssh
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias... Hecho
Leyendo la información de estado... Hecho
Se instalarán los siguientes paquetes adicionales:
  ncurses-term openssh-server openssh-sftp-server ssh-import-id
Paquetes sugeridos:
  molly-guard monkeysphere ssh-askpass
Se instalarán los siguientes paquetes NUEVOS:
  ncurses-term openssh-server openssh-sftp-server ssh ssh-import-id
0 actualizados, 5 nuevos se instalarán, 0 para eliminar y 24 no actualizados.
Se necesita descargar 836 kB de archivos.
Se utilizarán 6.804 kB de espacio de disco adicional después de esta operación.
¿Desea continuar? [S/n] S
Des:1 http://es.archive.ubuntu.com/ubuntu noble-updates/main amd64 openssh-sftp-
server amd64 1:9.6p1-3ubuntu13.5 [37,3 kB]
Des:2 http://es.archive.ubuntu.com/ubuntu noble-updates/main amd64 openssh-serve
r amd64 1:9.6p1-3ubuntu13.5 [509 kB]
Des:3 http://es.archive.ubuntu.com/ubuntu noble/main amd64 ncurses-term all 6.4+
20240113-1ubuntu2 [275 kB]
Des:4 http://es.archive.ubuntu.com/ubuntu noble-updates/main amd64 ssh all 1:9.6
p1-3ubuntu13.5 [4.658 B]
```

Figura Anexo 8. Instalación ssh

Después descargamos Hadoop de la página oficial y descomprimimos el archivo con el siguiente comando:

> tar -zxvf ~/Downloads/hadoop-3.2.3.tar.gz

A continuación, lo extraemos y lo movemos a la carpeta opt

Posteriormente debemos realizar algunas modificaciones en los ficheros de configuración que se muestran a continuación las cuales aparecen en la propia documentación de Hadoop:

```
<configuration>
 <property>
 <name>fs.defaultFS</name>
 <value>hdfs://localhost:9000</value>  </property>
 <property>
<name>hadoop.proxyuser.dataflair.groups</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.dataflair.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.groups</name> <value>*</value>
 </property>
</configuration>
```

Figura Anexo 9. Fichero core-site.xml

```
<configuration>
 <property>
 <name>dfs.replication</name>
 <value>1</value>
 </property>
</configuration>
```

Figura Anexo 10. Fichero hdfs-site.xml

```
<configuration>
 <property>
 <name>mapreduce.framework.name</name>  <value>yarn</value>
 </property>
 <property>
 <name>mapreduce.application.classpath</name>

<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/h
 </property>
</configuration>
```

Figura Anexo 11. Fichero mapred-site.xml

```
<configuration>
 <property>
 <name>yarn.nodemanager.aux-services</name>
 <value>mapreduce_shuffle</value>
 </property>
 <property>
 <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_
 </property>
</configuration>
```

Figura Anexo 12. Fichero yarn-site.xml

El siguiente paso es establecer una conexión SSH con el equipo, para ello ponemos el siguiente comando en la terminal: ssh localhost

Esto deberemos hacerlo cada vez que queramos ejecutar Hadoop.

También deberemos ejecutar las claves y cambiar los permisos.

```
asa117@alejandro:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
Generating public/private rsa key pair.
Your identification has been saved in /home/asa117/.ssh/id_rsa
Your public key has been saved in /home/asa117/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:CcCk9RdtkQx4VHJqQvlhTJLOl4JM1ktAW7LADHd+UKs asa117@alejandro
The key's randomart image is:
+---[RSA 3072]----+
|.+o+X+=B*=+o     |
| .o*+X*o===      |
|  .++=*=++       |
|    oo==+.       |
|    E  oS        |
|                 |
|                 |
|                 |
|                 |
+----[SHA256]-----+
```

Figura Anexo 13. Generación clave

Cuando ya están los ficheros modificados y todo configurado, procedemos a ejecutarlo y a comprobar que funcione correctamente para ello introducimos en la terminal el comando start-all.sh.

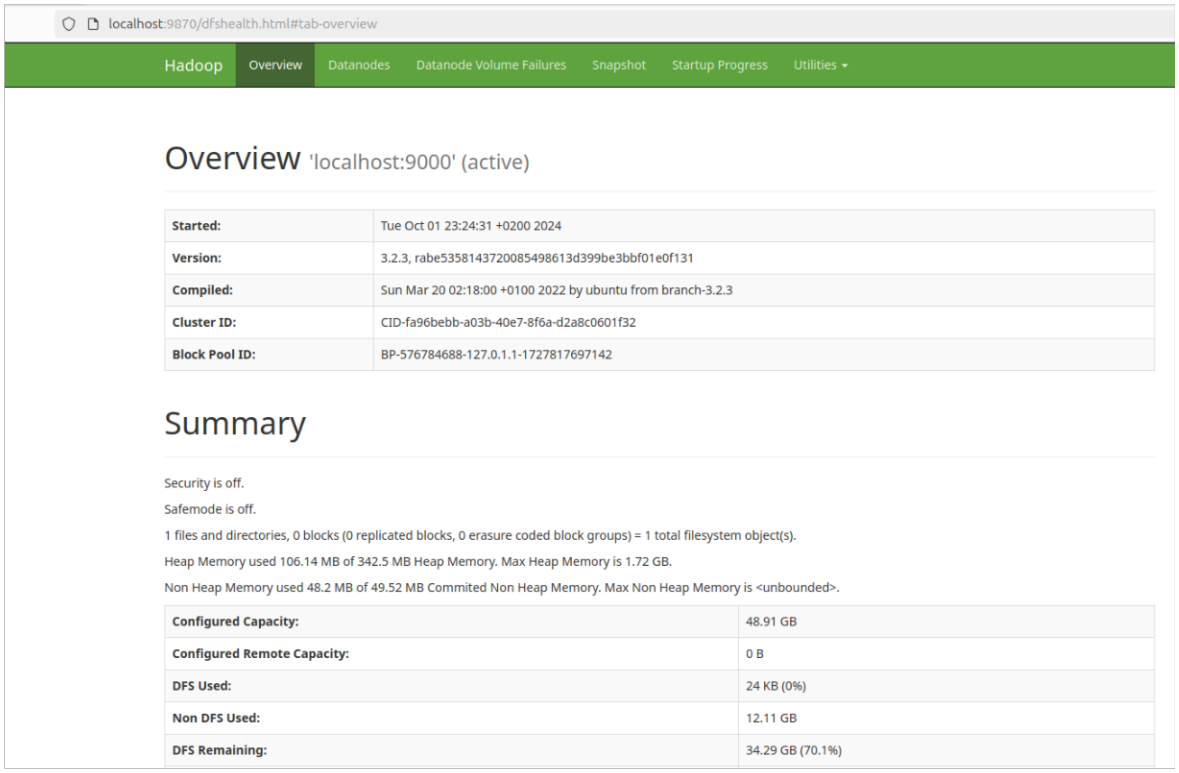


Figura Anexo 14. Hadoop activado

3. Apache Hive

En primer lugar, descargamos hive con los siguientes comandos:

wget https://dlcdn.apache.org/hive/hive-3.1.3/apache-hive-3.1.3-bin.tar.gz

tar -xvzf apache-hive-3.1.3-bin.tar.gz

Para instalar Apache Hive lo descargamos de la página oficial y después procedemos a instalar en nuestro equipo mysql-server.

```
asa117@alejandro:~$ sudo apt install mysql-server
[sudo] contraseña para asa117:
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias... Hecho
Leyendo la información de estado... Hecho
Se instalarán los siguientes paquetes adicionales:
  libaio1t64 libcgi-fast-perl libcgi-pm-perl libevent-core-2.1-7t64
  libevent-pthreads-2.1-7t64 libfcgi-bin libfcgi-perl libfcgi0t64
  libhtml-template-perl libmecab2 libprotobuf-lite32t64 mecab-ipadic
  mecab-ipadic-utf8 mecab-utils mysql-client-8.0 mysql-client-core-8.0
  mysql-common mysql-server-8.0 mysql-server-core-8.0
Paquetes sugeridos:
  libipc-sharedcache-perl mailx tinyca
Se instalarán los siguientes paquetes NUEVOS:
  libaio1t64 libcgi-fast-perl libcgi-pm-perl libevent-core-2.1-7t64
  libevent-pthreads-2.1-7t64 libfcgi-bin libfcgi-perl libfcgi0t64
  libhtml-template-perl libmecab2 libprotobuf-lite32t64 mecab-ipadic
  mecab-ipadic-utf8 mecab-utils mysql-client-8.0 mysql-client-core-8.0
  mysql-common mysql-server mysql-server-8.0 mysql-server-core-8.0
0 actualizados, 20 nuevos se instalarán, 0 para eliminar y 24 no actualizados.
Se necesita descargar 29,3 MB de archivos.
Se utilizarán 242 MB de espacio de disco adicional después de esta operación.
¿Desea continuar? [S/n] S
```

Figura Anexo 15. Instalación mysql-server

Para comprobar su funcionamiento ejecutamos en la terminal el siguiente comando:

```
asa117@alejandro:~$ sudo mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 8
Server version: 8.0.39-0ubuntu0.24.04.2 (Ubuntu)

Copyright (c) 2000, 2024, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
```

Figura Anexo 16. Inicio mysql-server

El siguiente paso es configurar la seguridad de mysql:

```
asa117@alejandro:~$ sudo mysql_secure_installation

Securing the MySQL server deployment.

Connecting to MySQL using a blank password.
The 'validate_password' component is installed on the server.
The subsequent steps will run with the existing configuration
of the component.

Skipping password set for root as authentication with auth_socket is used by def
ault.
If you would like to use password authentication instead, this can be done with
the "ALTER_USER" command.
See https://dev.mysql.com/doc/refman/8.0/en/alter-user.html#alter-user-password-
management for more information.

By default, a MySQL installation has an anonymous user,
allowing anyone to log into MySQL without having to have
a user account created for them. This is intended only for
testing, and to make the installation go a bit smoother.
You should remove them before moving into a production
environment.
```

Figura Anexo 17. Privilegios mysql-server

A continuación, procedemos a descargar mysql y lo guardaremos en la carpeta que este Hive para que pueda utilizado por Hive.

```
root@alejandro:/# mv ../home/asa117/Descargas/mysql-connector-j-8.0.33.jar opt/hive/lib/
```

Figura Anexo 18. Cambiar ubicación mysql

Después deberemos cambiar algunos ficheros de configuración:

En primer lugar, en el fichero .bashrc realizamos añadimos las siguientes líneas como root:

```
export HIVE_HOME=/opt/hive/
export PATH=$PATH:$HIVE_HOME/bin
```

Figura Anexo 19. Fichero .bashrc

```
export HADOOP_HOME=~/hadoop-3.2.3/
export HADOOP_HEAPSIZE=512
export HIVE_CONF_DIR=/opt/hive/conf
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export YARN_HOME=$HADOOP_YARN_HOME
```

Figura Anexo 20. Fichero hive-env.sh

```
<property>
<name>system:java.io.tmpdir</name>
<value>/tmp/hive/java</value>
</property>
<property>
<name>system:user.name</name>
<value>${user.name}</value>
</property>
```

Figura Anexo 21. Fichero hive-site.xml

Posteriormente inicializaremos un esquema de base de datos MySQL

Una vez modificado los ficheros empezamos a ejecutar un esquema de datos MySQL.

Por último, comprobamos el funcionamiento de Hive, cabe destacar que Hadoop debe estar ejecutándose.

```
asa117@alejandro:/usr/lib/hive/apache-hive-3.1.3-bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hive/apache-hive-3.1.3-bin/lib/log4j-
slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/asa117/hadoop-3.2.3/share/hadoop/common/
lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 2e47dbb6-2229-4bcc-a94c-4e2a56f35209

Logging initialized using configuration in jar:file:/usr/lib/hive/apache-hive-3.
1.3-bin/lib/hive-common-3.1.3.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
hive>
```

Figura Anexo 22. Hive en funcionamiento

4. Apache Zeppelin

A continuación, procedemos a la instalación de Apache Zeppelin. Una herramienta de desarrollo con la cual podemos crear nuestro código y ejecutarlo con las herramientas nombradas anteriormente.

En mi caso utilizaré la versión 0.11.0, ya que las versiones posteriores pueden dar problemas de compatibilidad con Spark.

Una vez descargado debemos cambiar el puerto para que no coincida con el de Spark. En mi caso pondré el 8880 y cambiamos el nombre a los ficheros de configuración.

```
<property>
  <name>zeppelin.server.port</name>
  <value>8880</value>
  <description>Server port.</description>
</property>
```

Figura Anexo 23. Fichero zeppelin-site.xml

```
114
115   export ZEPPELIN_PORT = 8880
116   # export JAVA_HOME=
117   export USE_HADOOP=~/hadoop-3.2.3
118   export SPARK_MASTER=/opt/spark
119   export SPARK_HOME=/opt/spark
120
```

Figura Anexo 24. Cambio de puerto zeppelin-env

Una vez realizados los cambios ya podremos iniciar Zeppelin. A continuación, mostramos un ejemplo de Zeppelin usando Spark. Cabe destacar que debemos configurar en modo standalone.

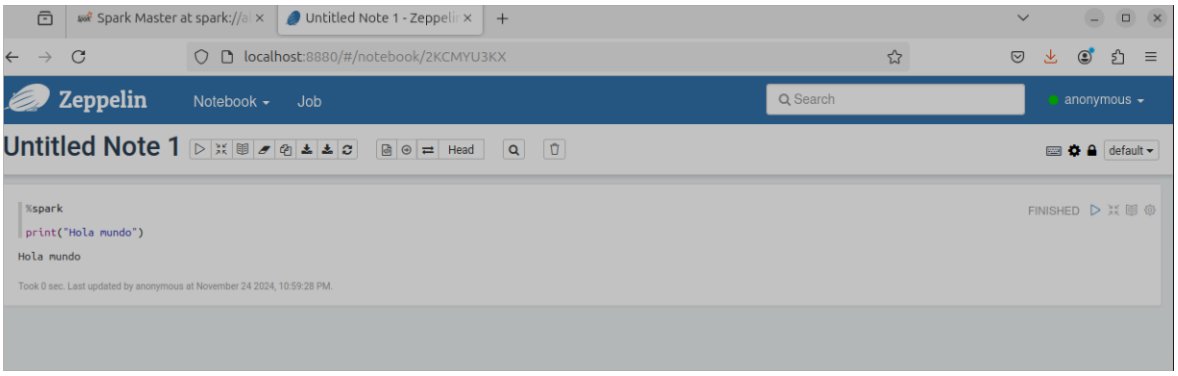


Figura Anexo 25. Zeppelin funcionando con Spark.

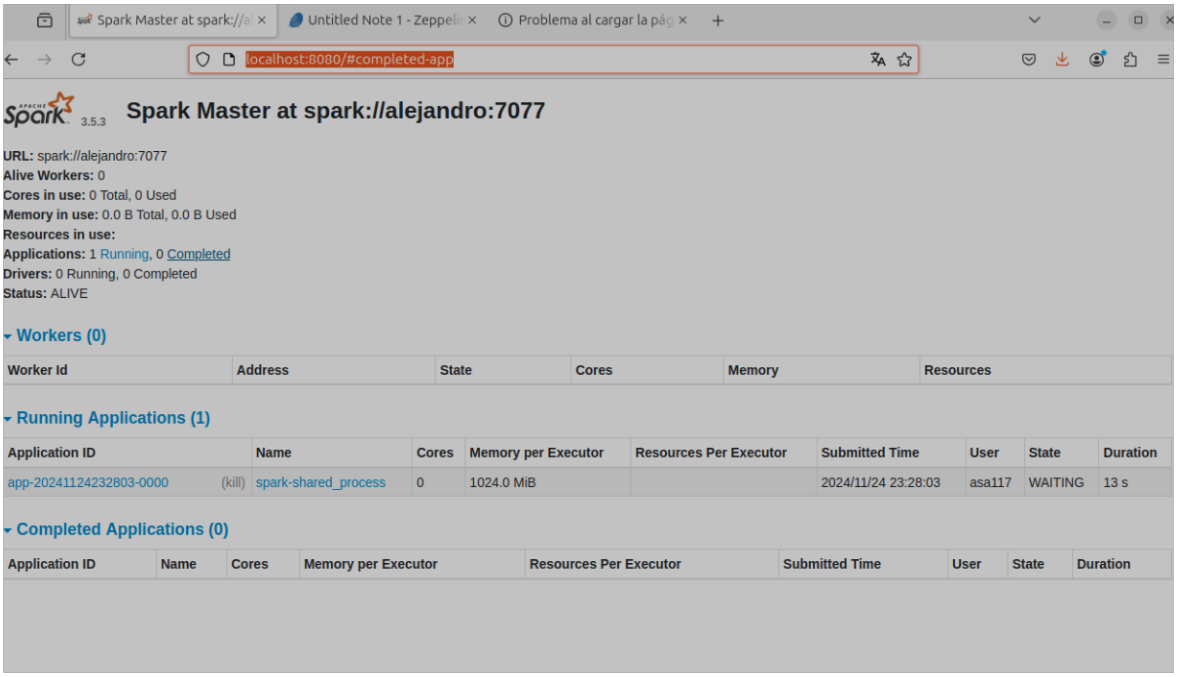


Figura Anexo 26. Proceso Zeppelin en Spark