یک servey مناسب درباره Hadoop و Sparkآماده نموده و تفاوت ها و شباهت ها و دامنه کاربرد هر یک را مشخص کنید . این پژوهش میبایست شامل جداول مقایسه ای و بیان دقیق موارد باشد .

در ابتدا قصد داریم هرکدام را به طور دقیق معرفی کنیم و بعد به تفاوت ها و شباهت ها بپردازیم :

SPARK : وقتی داده های خیلی بزرگ داشته باشیم برای پردازش این داده های خیلی بزرگ می توانیم از اسپارک استفاده کنیم, اسپارک مثل یه موتور خیلی قوی می تواند این داده های خیلی بزرگ را پردازش کند . ما اینجا درباره داده های خیلی بزرگ صحبت کردیم و داده های بزرگ به این معنی است که این داده هارا نمی توان روی یک سیستم پردازش کرد . این داده های بزرگ مثل داده هایی که حجم آن ها خیلی بیشتر از RAM ما هستند و این داده های خیلی بزرگ به صورت توزیع شده هستند , توزیع شده به این معنی است که این داده ها روی یک سیستم قرار ندارند بلکه روی چند سیستم هستند و به صورت جداگانه پردازش می شوند .

مثلا ما می خواهیم روی یک فایل بزرگ متنی مثل دفتردوم مثنوی عملیاتی انجام دهیم و ما در اینجا کلی حروف داریم که باید بررسی شوند همانطور که میدانید حجم این فایل خیلی زیاد است و قطعاً فیلتر کردن این فایل‌ سنگین،‌ به زمان زیادی احتیاج دارد. ممکن است مثلا در حدود یک روز پردازش روی این داده ها طول بکشد تا بتوانیم تمام حروف را پردازش و حروف های خاص را فیلتر کنیم .

برای حل این مشکل اسپارک یک راه حل بسیار مناسب است . به اینصورت که با استفاده از اسپارک می توانیم این پردازش ها را به صورت موازی و همزمان روی چند تا سیستم اجرا کنیم و نتیجه نهایی رو جمع آوری کنیم و روی یک سیستم نمایش بدیم .

: Hadoop در اینجا هم ما با داده های بسیار بزرگ سروکار داریم و برای توسعه و ذخیره سازی این داده های بزرگ از هدوپ استفاده می کنیم . هدوپ نیز مانند اسپارک روی داده های بزرگ روی چندین سیستم مختلف پردازش انجام می دهد و نتیجه نهایی را روی یک کامپیوتر که همان کامپیوتر مقصد است نمایش می دهد .

برای مثال ما یک سایت داریم که تعداد زیادی کاربر مثلا در حدود ۲۰ میلیون در آن ثبت شده است و نیاز است به ازای ورود آن ها نام و اطلاعاتی در یک فایل متنی ذخیر شود , پس یعنی ما برای هر کاربر یه سطر باید به فایلمون اضافه کنیم و حالا در نظر بگیرید که بخوایم اطلاعات یک کاربر را بررسی کنیم از بین این همه اطلاعات بدست آوردن اطلاعات نهایی خیلی زمان بر خواهد بود و برای این اطلاعات کلی باید منتظر ماند .

برای حل این مشکل میتوان از هدوپ استفاده کرد. در اینجا هدوپ، فایل متنی را در یک سیستم بین چندین سیستم پردازش و بعد ذخیره میکند. پردازش، به صورت موازی و همزمان، بر روی هر کدام از سیستم ها، انجام می شود، و در نهایت، توسط یک سیستم، همه ی اطلاعات پردازش شده، جمع آوری می شود و نتیجه نهایی اعلام می شود .

حال به کاربردهای spark و Hadoopمی پردازیم :

کاربردهای spark :

امروزه شرکت های برتر مانند علی بابا، یاهو، اپل، گوگل، فیس بوک و نتفلیکس از Spark استفاده می کنند.

Spark یک چارچوب محاسباتی خوشه ای منبع باز با قابلیت پردازش درون حافظه است. این در زبان برنامه نویسی اسکالا توسعه یافته است. در حالی که Spark شبیه MapReduce است، ویژگی ها و قابلیت های بیشتری دارد که آن را به یک ابزار کلان داده کارآمد تبدیل می کند. بسیاری از API های تعاملی را در چندین زبان از جمله Scala، Java، Python و R ارائه می دهد .

دلایل محبوبیت Spark :

* Spark مورد علاقه توسعه دهندگان است زیرا به آنها اجازه می دهد برنامه هایی را در جاوا، اسکالا، پایتون و حتی R بنویسند.
* Spark توسط یک جامعه توسعه دهندگان فعال پشتیبانی می شود، و همچنین توسط یک شرکت اختصاصی - Databricks پشتیبانی می شود.
* اگرچه اکثر برنامه‌های Spark از HDFS به عنوان لایه ذخیره‌سازی فایل داده استفاده می‌کنند، اما با سایر منابع داده مانند Cassandra، MySQL و AWS S3 نیز سازگار است.
* Spark در بالای اکوسیستم Hadoop توسعه یافته است که امکان استقرار آسان و سریع Spark را فراهم می کند.
* Spark از یک فن‌آوری خاص، اکنون به یک فناوری اصلی تبدیل شده است، به لطف انبوهی از داده‌های تولید شده توسط تعداد رو به رشد اینترنت اشیا و سایر دستگاه‌های متصل.

کاربردهای آپاچی اسپارک :

از آنجایی که پذیرش Spark در صنایع به طور پیوسته در حال افزایش است، کاربردهای Spark منحصر به فرد و متنوعی را به وجود آورده است.

برنامه های Spark :

1. پردازش داده های جریانی :

شگفت انگیزترین جنبه Apache Spark توانایی آن در پردازش داده های جریان است. در هر ثانیه، حجم بی سابقه ای از داده ها در سطح جهان تولید می شود. این امر شرکت‌ها و کسب‌وکارها را به پردازش داده‌ها در حجم زیاد و تجزیه و تحلیل آن‌ها در زمان واقعی وا می‌دارد. ویژگی Spark Streaming می تواند به طور موثر این عملکرد را مدیریت کند. Spark Streaming با یکی کردن قابلیت‌های پردازش داده‌های متفاوت، به توسعه‌دهندگان این امکان را می‌دهد تا از یک چارچوب واحد برای تطبیق با تمام نیازهای پردازشی خود استفاده کنند.

برخی از بهترین ویژگی های Spark Streaming عبارتند از:

– Spark’s Streaming ETL آنها ابتدا داده ها را می خوانند، سپس آن را به فرمت سازگار با پایگاه داده تبدیل کنید و در نهایت آن را در پایگاه داده هدف بنویسید.

غنی‌سازی داده – این ویژگی با ترکیب آن با داده‌های استاتیک به غنی‌سازی کیفیت داده‌ها کمک می‌کند، بنابراین، تجزیه و تحلیل داده‌های بلادرنگ را ارتقا می‌دهد.

شناسایی رویداد ماشه - ویژگی تشخیص رویداد ماشه به شما امکان می دهد رفتارهای غیرعادی یا "رویدادهای ماشه ای" را که می توانند سیستم را به خطر بیندازند یا مشکل جدی در آن ایجاد کنند، به سرعت شناسایی کرده و به آنها پاسخ دهید.

تجزیه و تحلیل جلسه پیچیده – به شما امکان می دهد جلسات و رویدادهای زنده (به عنوان مثال، فعالیت کاربر پس از ورود به وب سایت/برنامه) را با هم گروه بندی کنید و همچنین آنها را تجزیه و تحلیل کنید. علاوه بر این، از این اطلاعات می توان برای به روز رسانی مدل های ML به طور مداوم استفاده کرد. نتفلیکس از این ویژگی برای به دست آوردن بینش رفتار مشتری در زمان واقعی در پلتفرم و ایجاد توصیه های نمایش هدفمندتر برای کاربران استفاده می کند.

2. یادگیری ماشینی

اسپارک توانایی های ستودنی در یادگیری ماشینی دارد. این به یک چارچوب یکپارچه برای انجام تجزیه و تحلیل پیشرفته مجهز است که به شما امکان می دهد پرس و جوهای مکرر را روی مجموعه داده ها اجرا کنید. این، در اصل، پردازش الگوریتم های یادگیری ماشینی است. کتابخانه یادگیری ماشین (MLlib) یکی از قوی ترین اجزای ML اسپارک است.

این کتابخانه می تواند خوشه بندی، طبقه بندی، کاهش ابعاد و موارد دیگر را انجام دهد. با MLlib، Spark را می توان برای بسیاری از عملکردهای Big Data مانند تجزیه و تحلیل احساسات، هوش پیش بینی، تقسیم بندی مشتری و موتورهای توصیه و موارد دیگر مورد استفاده قرار داد.

یکی دیگر از کاربردهای قابل ذکر Spark امنیت شبکه است. با استفاده از مولفه‌های متنوع پشته Spark، ارائه‌دهندگان/شرکت‌های امنیتی می‌توانند بسته‌های داده را در زمان واقعی بازرسی کنند تا هر گونه اثری از فعالیت مخرب را شناسایی کنند.

3. محاسبات مه

درک مفهوم محاسبات مه عمیقاً با اینترنت اشیا در هم تنیده است. اینترنت اشیا بر اساس ایده جاسازی اشیا و دستگاه‌هایی با حسگرهایی که می‌توانند بین یکدیگر و همچنین با کاربر ارتباط برقرار کنند، رشد می‌کند، بنابراین، شبکه‌ای به هم پیوسته از دستگاه‌ها و کاربران ایجاد می‌کند. همانطور که کاربران بیشتر و بیشتر از پلتفرم های اینترنت اشیا استفاده می کنند و کاربران بیشتری به وب دستگاه های متصل به هم می پیوندند، میزان داده های تولید شده فراتر از درک است.

دلایل محبوبیت Hadoop :

Apache Hadoop یک چارچوب منبع باز Big Data است که برای ذخیره و پردازش Big Data و همچنین برای توسعه برنامه های کاربردی پردازش داده در یک محیط محاسباتی توزیع شده استفاده می شود. برنامه های کاربردی مبتنی بر Hadoop بر روی مجموعه داده های بزرگی اجرا می شوند که در خوشه هایی از رایانه های کالایی که ارزان و ارزان هستند پخش شده اند. بنابراین، شما قدرت محاسباتی یک شبکه خوشه ای گسترده را با هزینه ای مقرون به صرفه دریافت می کنید. ساختار سیستم فایل توزیع شده Hadoop امکان پردازش همزمان و تحمل خطا را فراهم می کند.

ویژگی های Hadoop :

۱. این برای تجزیه و تحلیل داده های بزرگ مناسب است

به طور معمول، داده های بزرگ ماهیتی بدون ساختار و توزیع شده دارند. این همان چیزی است که خوشه های Hadoop را برای تجزیه و تحلیل داده های بزرگ مناسب می کند. Hadoop به جای داده های واقعی، منطق پردازش به گره های محاسباتی جریان می یابد و در نتیجه پهنای باند شبکه کمتری مصرف می شود. این باعث افزایش کارایی برنامه های Hadoop می شود.

۲. مقیاس پذیر است

بهترین چیز در مورد خوشه‌های Hadoop این است که می‌توانید با افزودن گره‌های خوشه‌ای اضافی به شبکه، بدون اعمال هیچ گونه تغییری در منطق برنامه، آن‌ها را به هر میزانی مقیاس دهید. بنابراین، با افزایش حجم، تنوع و سرعت Big Data، می‌توانید خوشه Hadoop را برای برآورده کردن نیازهای داده‌ای در حال رشد، مقیاس‌بندی کنید.

۳. عیب‌پذیر است

در Hadoop، مقرراتی برای تکثیر داده های ورودی به سایر گره های خوشه ای نیز وجود دارد. بنابراین، اگر یک گره خوشه ای از کار بیفتد، پردازش داده متوقف نخواهد شد زیرا گره خوشه دیگری می تواند جایگزین گره شکست خورده شود و روند را ادامه دهد.

برنامه های کاربردی Hadoop :

امنیت و اجرای قانون :

Hadoop در حال حاضر به عنوان یک ابزار فعال در اجرای قانون استفاده می شود. به لطف تجزیه و تحلیل سریع و قابل اعتماد Big Data، Hadoop به سازمان های مجری قانون (مانند اداره پلیس) کمک می کند تا فعال تر، کارآمدتر و پاسخگوتر شوند.

افزایش رضایت مشتری و نظارت بر شهرت آنلاین :

کسب‌وکارها در حال حاضر از Hadoop برای تجزیه و تحلیل داده‌های فروش و مقایسه آن با بسیاری از عوامل دیگر استفاده می‌کنند تا مشخص کنند که یک محصول خاص چه زمانی و در چه زمانی بهترین فروش را دارد. با نظارت مستمر بر داده‌های فروش، صاحبان مشاغل می‌توانند دریابند که چرا محصولات خاص در روزها یا ساعت‌ها یا فصل‌های خاص بهتر می‌فروشند. به همین ترتیب، Hadoop همچنین می‌تواند رسانه‌های اجتماعی و مکالمات آنلاین را استخراج کند تا ببیند مشتریان شما (اعم از موجود و بالقوه) در مورد شما در پلتفرم‌های آنلاین چه می‌گویند. احساسات پشت نظرات و بازخورد مشتریان را زیر نظر دارد.

وضعیت حیاتی بیمار را پایش کنید :

بسیاری از بیمارستان ها استفاده از Hadoop را برای بهره وری بیشتر کارکنان خود در فرآیند کاری خود آغاز کرده اند. سیستم‌ها و ماشین‌های بهداشتی حجم زیادی از داده‌های بدون ساختار تولید می‌کنند. سیستم های پردازش داده های مرسوم نمی توانند چنین مقادیر زیادی از داده های خام را پردازش و تجزیه و تحلیل کنند .

تفاوت های spark و Hadoop :

|  |  |
| --- | --- |
| SPARK | HADOOP |
| موتور سریع و عمومی برای پردازش داده های در مقیاس بزرگ | نرم افزار ایست که میتواند از یک سیستم محاسباتی منفرد تا نزدیک به هزاران سیستم کالایی که به آنها امکان ذخیره سازی محلی و قدرت رایانه را می دهد |
| پردازش در حافظه سریع تر | پردازش در حافظه کندتر |
| پرس و جوهای تعاملی ، بار کاری و از همه مهمتر یادگیری ماشین بهتر | ضعیف بودن پرس و جوهای تعاملی ، بار کاری و یادگیری ماشین |
| اسپارک برای مدیریت پرونده وابسته است به هدوپ | به صورت مستقل عمل می کند |
| ضعیف تر و قابلیت گسترش کمتر | قدرت مند تر و قابلیت گسترش بیشتر |
| سرعت بخشی کندتر به دلیل real time بودن | سرعت بخشی بالاتر به دلیل پردازش در حافظه |
| زبان های برنامه نویسی : java , scala , python | زبان های برنامه نویسی : java |
| سهولت استفاده و تعداد عملکرد های بالا در API | دارای فضای مکانی بسیار بالا |
| برای پردازش داده هاست | دارای زیرساخت های توزیع شده |
| در برابر خرابی ها مقاوم نیست | در برابر خرابی ها مقاوم تر است |
| اسپارک یک موتور پردازش داده است | هدوپ برای جستجوی متن هاست |
| Spark یک زیر پروژه فریم ورک هدوپ است | هدوپ یک framework است |
| اسپارک از پردازش همزمان و موازی استفاده می کند | هدوپ از پردازش گروهی استفاده می کند |
| اسپارک از memory استفاده می کند | هدوپ برای ذخیره و پردازش از دیسک استفاده می کند |

شباهت های spark و Hadoop :

|  |
| --- |
| هردو برای پردازش big Data هستند |
| انجام توابع به صورت موازی |
| هردو متقابلا منحصر به فرد نیستند |
| نمی توانند جایگزینی کامل برای یکدیگر باشند |
| کاملا با هم سازگار هستند |
| تیم ها در هر دو راه حل های بسیار قدرتمندی برای حل داده های بزرگ ارائه می دهند |
| اجرا کردن  کارکردهای یکسان |
| بازیابی فاجعه در هر دو فناوری به خوبی پیاده سازی شده است |
| مکمل یک دیگر هستند |
| هردو open source و هزینه ایی ندارند |