Database Systems

ایجاد بانک های اطلاعاتی

مدرس: ابوالقاسم حسن پور

مطالعه ی بسیار و پی گیر در مسائل علمی ، باعث شگفتی عقل و تقویت نیروی فکر و فهم است. امام صادق (ع)

فصل پنجم

داده های بزرگ (Big Data)

معرفی

معرفی

- Big Data یا کلان‌داده معمولاً شامل مجموعه اطلاعاتی است که بزرگ، متنوع با ساختاری پیچیده و دارای دشواری هایی برای ذخیره سازی ، تحلیل و تصویر سازی ، پردازش های بیشتر یا نتایج می باشد.
- داده های بیشتر نیازمند تحلیل های دقیق تری است. تحلیل های دقیق تر منجر به تصمیم گیری های مطمئن بیشتری شده و تصمیمات بهتر، می تواند معنای کارایی بیشتر عملیات، کاهش هزینه ها و کاهش ریسک ها باشد.

Kilo → Mega → Giga → Tera → Peta → Exa → Zetta → Yotta

معرفی

- Big Data از تراکنش های ایمیل ها، ویدئوها، صوت ها، کلیک کردن ها، Logها و ارسال ها، درخواست های جستجو، یادداشت ها، تعاملات شبکه های اجتماعی، داده های علمی ، سنسورها، تلفن ها و برنامه های کاربردی آنها تولید می شوند. این داده ها به صورت نمایی رشد می کنند و ذخیره می شوند.
- ذخیره سازی، مدیریت، به اشتراک گذاری، تحلیل و نمایش آنها از طریق ابزارهای نوعی پایگاه داده ها دشوار است.

Big Data
VVV

چالش‌ها و فرصتهای توسعه اطلاعات دارای سه بعد می‌باشد:

- Volume: حجم اطلاعات (مقدار اطلاعات)
- Velocity: سرعت (سرعت اطلاعات خروجی و ورودی)
- Variety: تنوع (دامنه نوع اطلاعات و منابع)

Big Data
Volume

- حجم: فاکتورهای بسیاری به افزایش حجم داده ها کمک می کند.
  - داده های بر پایه تراکنش ذخیره شده در طی سالیان
  - داده های غیرساختارمند سرازیر شده از رسانه های اجتماعی
  - مقدار در حال افزایش داده های ماشین-به-ماشین و سنسور جمع آوری شده
- در گذشته، حجم انبوه داده یک مسئله ذخیره کردن بود. اما با کاهش هزینه های ذخیره، مسائل دیگری سر بر می آورند؛ شامل چگونگی تعیین ارتباط در حجم زیاد داده ها و چگونگی استفاده از علم تجزیه و تحلیل به منظور ایجاد ارزش از داده - های مرتبط.

Big Data Velocity

- سرعت: داده ها با سرعتی بی سابقه وارد شده و باید در زمان مناسب به سراغ آن‌ها رفت.
- تگ های RFID، سنسورها و اندازه گیری هوشمند، نیاز به سر و کله زدن با جریانات داده را در اولین زمان نزدیک به اکنون را ایجاد می کنند.
- واکنش سریع به کار با سرعت داده ها، چالشی برای بیشتر سازمان هاست.

Big Data Variety

- تنوع: داده ها به شکل های گوناگونی وارد می شوند.
- داده های عددی ساختاریافته در پایگاه های داده سنتی
- اطلاعات ایجاد شده از برنامه های کاربردی کسب وکار
- اسناد متنی غیرساختاریافته، ایمیل، صدا و تراکنش های مالی
- مدیریت، ادغام و حاکمیت بر انواع گوناگون داده، چیزی است که بسیاری از سازمان ها هنوز با آن درگیرند.

ویژگی های Big Data

- حجم و اندازه: اندازه داده‌های تولید شده و ذخیره شده. اندازه‌ی داده در شناسایی ارزش یا کلانگی داده کلیدی است. اگر داده خُرد باشد، کلان داده خوانده نمی‌شود.
- تنوع: نوع و ماهیت داده. دسته‌بندی داده‌ها به گونه‌ها به شناخت بهتر می‌انجامد.
- سرعت: همان سرعت تولید داده است. نرخ بالای تولید داده، چالش‌هایی را در زمینه‌ی ذخیره‌سازی و پردازش داده پدید می‌آورد.
- تغییر پذیری: ناسازگاری داده می تواند پردازش‌ها را از رسیدگی و مدیریت داده بازدارد.
- صحت: کیفیت داده‌ی گردآوری شده می‌تواند بر تجزیه و تحلیل دقیق داده اثر بگذارد.

نمونه هایی از Big Data

- بهداشت و درمان:

سیستمهای پشتیبانی تصمیم گیری بالینی ، تجزیه و تحلیل فردی به کار برده شده برای مشخصات بیمار ، پزشکی شخصی ، عملکرد مبتنی بر ارزشگذاری برای پرسنل ، تحلیل الگوهای بیماری ، بهبود سلامت عمومی.

- بخش عمومی:

ایجاد شفافیت به واسطه داده های وابسته در دسترس ، کشف نیازها ، بهبود عملکرد ، اقدامات سفارشی برای محصولات مناسب و خدمات ،تصمیم گیری با سیستمهای اتوماتیک برای کاهش ریسکها ، نوآوری در محصولات جدید و خدمات.

نمونه هایی از Big Data

- جزئی:

  در تحلیل رفتار ذخیره سازی ، بهینه سازی قیمت و تنوع ، طراحی تبلیغ محصول ، توسعه عملکرد ، بهینه سازی ورودی کار ، بهینه سازی تدارکات و توزیع ، بازارهای مبتنی بر web

- ساخت:

  توسعه پیش بینی تقاضا ، برنامه ریزی زنجیره تأمین(ذخیره) ، پشتیبانی فروش ، توسعه عملیات تولید ، برنامه های کاربردی مبتنی بر جستجو در . web

- داده های مکان های شخصی:

  مسیر یابی هوشمند ، تبلیغات جغرافیایی هدفمند یا واکنش های اضطراری ، برنامه ریزی شهری ، مدلهای کسب و کار جدید.

روش ها Map Reduce

- Map Reduce یک Framework برنامه نویسی است برای محاسبات توزیع شده که به وسیله Google تولید شده و از روش تقسیم و غلبه استفاده میکند جهت درهم شکستن مسائل داده های حجیم مختلط به بخشهای کاری کوچک و پردازش موازی آنها

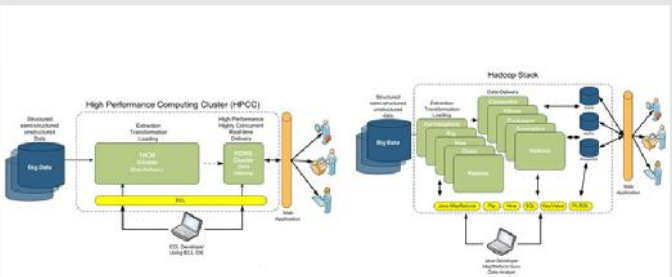
روش Map Reduce

- Map Reduce می تواند به دو مرحله تقسیم شود:
- Map Step: داده گره اصلی به تعدادی زیر مسئله کوچکتر خرد می شود. یک گره کارگر تعدادی زیر مجموعه از مسئله های کوچکتر را تحت کنترل گره دنبال کننده کار پردازش می کند و نتایج را در سیستم فایل محلی ذخیره می کند. جائیکه یک کاهنده قادر به دسترسی به آن باشد.
- Reduce Step: این مرحله داده های ورودی از مراحل نگاشت را تحلیل و ادغام می کند. میتواند چندین وظیفه کاهش جهت موازی سازی اجتماع ، وجود داشته باشد و این وظایف بر روی نودهای کارگر تحت کنترل دنبال کننده کار انجام میشود.

روش Hadoop

- Hadoop یک Framework مبتنی بر جاوا و سکوی متن باز ناهمگون است.
- Hadoop شامل یک سیستم فایل توزیع شده ، تجزیه و تحلیل و سکوی ذخیره سازی داده می باشد و یک لایه ای که محاسبات موازی، گردش کار و مدیریت پیکربندی را اداره میکند.
- HDFS: (Hadoop Distributed File System) یا سیستم فایل توزیع شده Hadoop، درمیان گره ها در یک خوشه Hadoop، اجرا می شود و سیستمهای فایل تعدادی داده ورودی و خروجی را به هم متصل میکند تا آنها را به صورت یک سیستم فایل بزرگ درست کند.

روش Hadoop



روش Hadoop

- HDFS: یک سیستم فایل توزیع شده بسیار تحمل کننده خطا است که مسئول ذخیره سازی داده ها در کلاسترها می باشد.
- MapReduce: یک تکنیک برنامه نویسی قدرتمند برای پردازش موازی کلاسترها است.
- Hbase: یک پایگاه داده توزیع شده مقیاس پذیر برای دسترسی خواندن/نوشتن به طور تصادفی است.
- Pig: یک سیستم پردازش داده سطح بالا برای تحلیل مجموعه های داده که به وسیله یک زبان سطح بالا رخ می دهد.
- Hive: یک برنامه کاربردی ذخیره سازی داده است که یک رابط (interfac) مشابه SQL و مدل رابطه ای را فراهم می آورد.

روش Hadoop

- Sqoop: یک پروژه برای انتقال داده بین پایگاه داده رابطه ای و Hadoop
- Avro: یک سیستم از داده های مرتب.
- Oozie: یک جریان کار برای کارهای Hadoop وابسته.
- Chukwa: یک زیر پروژه Hadoop به عنوان سیستم جمع آوری داده برای نظارت سیستم های توزیع شده.
- Flume: مجموعه logهای جاری توزیع شده و قابل اعتماد.
- Zookeeper: یک سرویس مرکزی است جهت فراهم آوردن همزمانی توزیع شده وسرویس های گروهی.