پژوهشکده بیمه

سادەسازی و تسریع خطوط لوله داده در برنامه‌های مالی دیجیتال و بیمه

مجری: الهام نوبری

.....


بسم الله الرحمن الرحيم

اطلاعات مجری و همکاران

مجری طرح	دکتر الهام نوبری استادیار دانشکده علوم دانشگاه علم و فناوری مازندران
مشاور طرح	دکتر امیرتیمور پاینده عضو هیات علمی دانشگاه شهید بهشتی
ناظر علمی طرح	دکتر فرزان خامسیان عضو هیات علمی پژوهشکده بیمه
تاریخ شروع اجرای طرح	پاییز ۱۴۰۱
تاریخ اتمام طرح	بهار ۱۴۰۲

بیان مسئله و ضرورت انجام طرح

- فعالیت‌های پیچیده روزانه در شرکتهای بیمه‌ای و مالی مانند محاسبه ریسک، بدهی یا اعتبار
- بزرگی حجم داده‌ها و زمان‌بری تجزیه و تحلیل داده‌ها بیشتر از زمان مجاز برای یک کسب و کار
- نکهداری داده‌های اخیر توسط پایگاه داده از نوع SQL و NoSQL فقط برای مدت کوتاه
- مشکلات استفاده از پایگاه داده‌های تحلیلی کندتر مانند انبارهای داده و نیز دریاچه‌های داده
- لزوم طراحی معماری مناسب پایگاه داده جهت فراهم کردن موارد زیر

صرفه جویی در هزینه فناوری اطلاعات

متناسب بودن هزینه عملیات و پشتیبانی با زمان پردازش
هرچه پردازش کوتاهتر باشد تیم‌های فنی کوچکتر
در پلتفرم‌های ابری هزینه‌ها متناسب با زمان پردازش

بهبود بهره وری

تقسیم کل زمان صرف شده بین زمانی که تحلیل‌گران داده و دانشمندان داده به صورت دستی با داده‌ها کار می‌کنند و زمانی که الگوریتم‌های کامپیوتری پردازش داده را انجام می‌دهند

کاهش خطرات نظارتی و عملیاتی

انجام فرآیندهای دسته‌ای روزانه، هفتگی و ماهانه در بازه زمانی محدود

ارائه قابلیت ها و خدمات جدید

استفاده‌ی شرکتهای مالی و بیمه‌ای از توانایی خود مبتنی بر داده‌ها برای پردازش اطلاعات بیشتردر زمان کمتر و در نتیجه ایجاد مزیت رقابتی و ارائه محصول بهتر

اهداف/سوالات/فرضیه های طرح

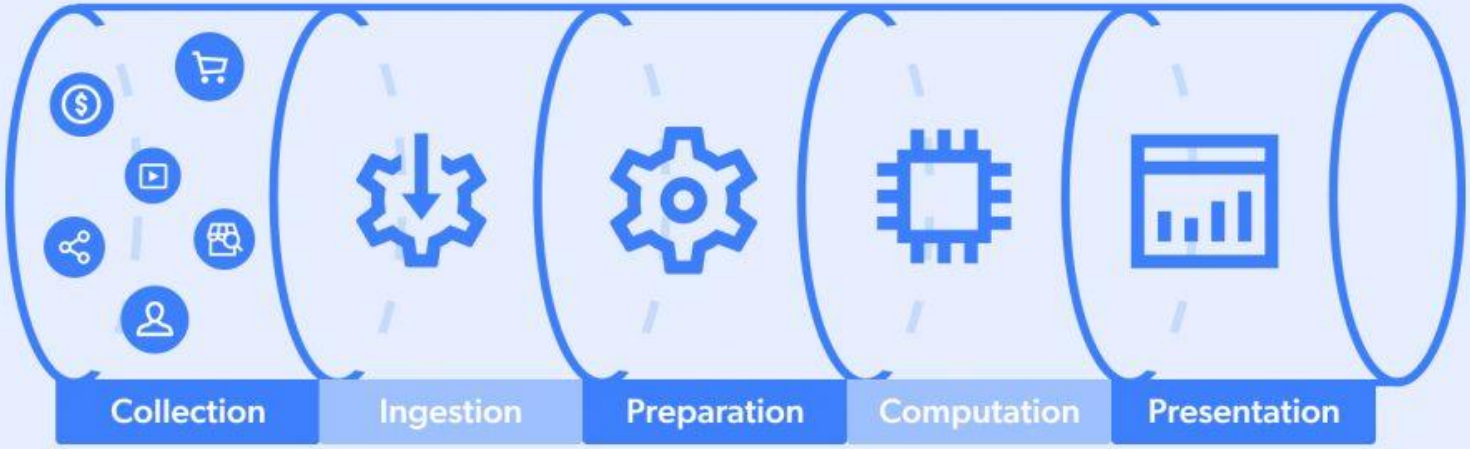
- خطوط لوله داده چیست؟
- خطوط لوله داده‌های بانکی و بیمه معمولاً از چه مراحلی تشکیل می شوند؟
- با توجه به اهمیت سرعت بخشیدن به خطوط لوله داده، برای تسریع چه کاری می توان انجام داد؟
- در پایگاه داده سنتی ذخیره داده دریافت داده، تبدیل داده ها و تولید خروجی مورد نیاز چگونه انجام می‌شود؟
- پایگاه داده LeanXcale چطور می‌تواند بهتر عمل کند؟

LeanXcale با این چشم‌انداز متولد شد تا فناوری پایگاه داده بزرگی در آینده باشد که شکاف بین دنیای داده‌های عملیاتی و دنیای داده‌های تحلیلی را پر کند. LeanXcale ریشه در تحقیقات فنی عمیق در سیستمهای توزیع شده و مدیریت داده دارد. اساتید آزمایشگاه سیستمهای توزیع شده تصمیم گرفتند تمام تحقیقات انجام شده برای بیش از ۱۵ سال تحقیق را کنار بگذارند و از ابتدا شروع کنند تا یک مدیر تراکنشی کاملاً متفاوت را طراحی کنند که بتواند بدون محدودیت مقیاس شود. بیش از نه ماه تحقیق برای تولید اولین نسخه از الگوریتم طول کشید، اما نتایج زیبا و ظریف بودند، راه حلی عالی که بزرگترین و مشکل سازترین گلوگاه در پایگاه های داده را برای چندین دهه حل کرد. سپس تیم تحقیقاتی به دنبال بودجه تحقیق و توسعه در اروپا بود، یک سری کمک هزینه دریافت کرد نمونه اولیه تحقیقاتی در سال ۲۰۱۰ تولید شد و تا سال ۲۰۱۵ تکمیل شد. در سال ۲۰۱۰ یک حق اختراع در مورد چگونگی مقیاس پردازش پرس و جو ثبت شد و در سال ۲۰۱۱ یک حق اختراع دیگر ثبت شد.

خطوط لوله داده چیست؟

خط لوله داده ها به مجموعه فرآیندهایی گفته می‌شود که برای جمع آوری، تبدیل، تحلیل و انتقال داده ها به کار می‌روند تا بتوان در قالبی مناسب از آن ها برای برنامه های کاربردی مختلف استفاده کرد. به طور معمول، برای ساخت این خطوط انتقال داده‌ها از ابزارها و فرآیندهایی استفاده می‌شود که در هر مرحله از جمع آوری و پردازش داده ها، داده ها را به صورت خودکار از یک مرحله به مرحله دیگر منتقل می کنند.

در یک Data Pipeline، داده ها معمولا از منابع مختلف مانند پایگاه های داده، فایل‌ها، سرویس های وب و سایر منابع جمع آوری می‌شوند. سپس، این داده ها به صورت خودکار در مراحل مختلفی مانند تبدیل، تحلیل، پاکسازی و ترکیب داده‌ها پردازش می شوند. در نهایت، داده‌ها به سیستم مورد نظر انتقال داده می شوند تا برای استفاده در برنامه های کاربردی و سایر سیستم‌ها موجود قرار گیرند. استفاده از Data Pipeline در سازمان‌ها، می تواند به بهبود کارایی و دقت در جمع آوری و پردازش داده‌ها و به کاهش زمان و هزینه مورد نیاز برای پردازش داده ها کمک می کند.


Collection
Ingestion
Preparation
Computation
Presentation

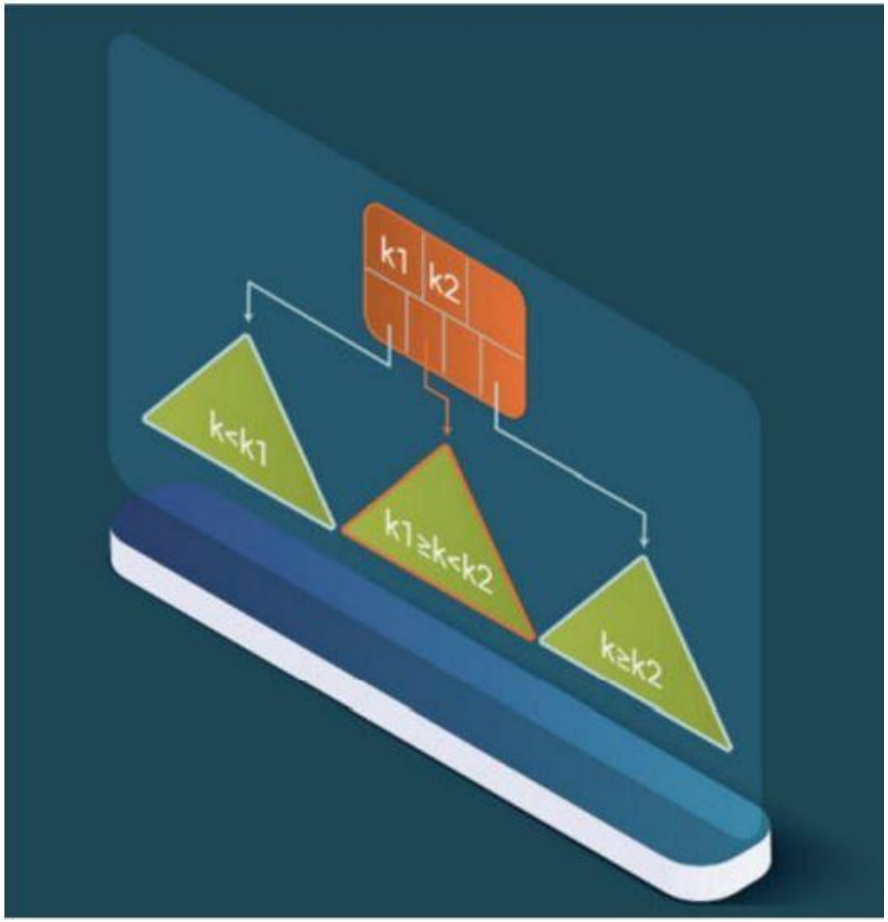
- دریافت داده

دریافت داده‌ها بسیار زیاد و صرف زمان قابل توجه

- چالش

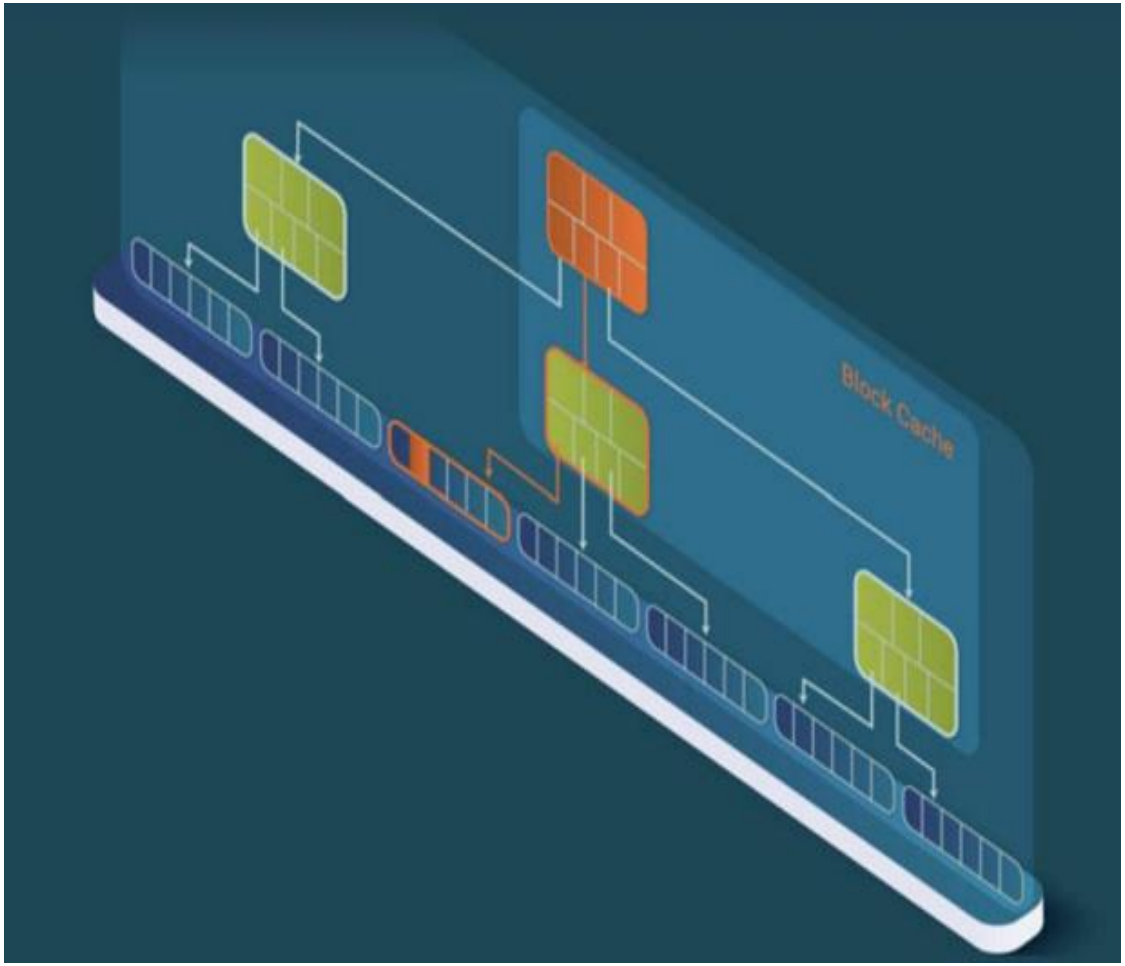
تسریع جذب داده‌ها بدون به خطر افتادن سرعت پرس وجو برای مرحله بعدی

راه حل؟؟؟؟



راه حل سنتی جذب داده در SQL

استفاده از ساختار داده پایدار به نام درخت B+ یافتن داده‌ها با پیچیدگی لگاریتمی

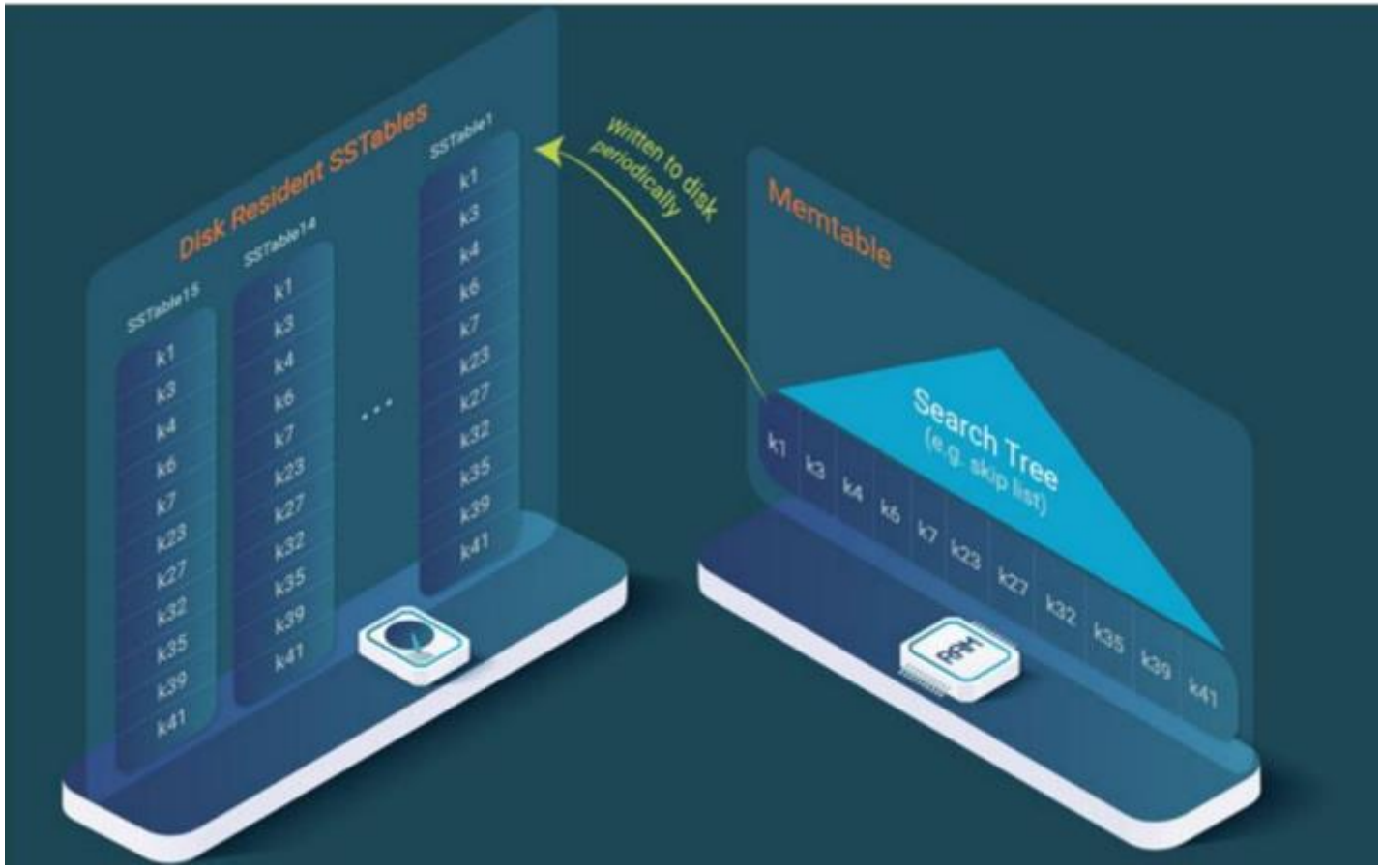


حافظه cash

با افزایش اندازه جدول، دریافت داده‌ها گران‌تر و گران‌تر می شود زیرا سطوح بیشتری در درخت B + به وجود می‌آید و هر ردیفهزینه بیشتری برای خواندن و خارج کردن از حافظه پنهان دارد

سریعترین پایگاه‌های داده NoSQL برای دریافت داده‌ها، از ذخیره‌سازی داده کلید-مقدار و پایگاه‌های داده ستونی استفاده می‌کنند. راه حل های مبتنی بر هش حتی از پرس و جوهای محدوده ساده پشتیبانی نمی‌کنند.

راه حل‌های دیگر مبتنی بر رویکرد جداول مرتب‌سازی شده رشته‌ای (SSTables) است که از آنچه در Google Big Table انجام می‌شود، تقلید می‌کنند.


Disk Resident SSTables
SSTable15
SSTable14
SSTable1
k1
k3
k4
k6
k7
k23
k27
k32
k35
k39
k41
Written to disk periodically
Memtable
Search Tree
(e.g. skip list)
k1
k3
k4
k6
k7
k23
k27
k32
k35
k39
k41
RAM

مرحله دوم خطوط لوله داده در بانکداری و بیمه

فرآیند تجزیه و تحلیل و غنی‌سازی داده‌هاست
تجمیع اطلاعات با استفاده از پرس‌وجوهای SQL

انبارهای داده در پاسخ به این نوع پرس و جوها خوب هستند اما از به روزرسانی‌ها پشتیبانی نمی‌کنند و فقط پرس و جوهای خواندنی را پشتیبانی می‌کنند

پایگاه های داده NoSQL

عدم پشتیبانی از داده‌های انبوه
عدم توانایی پس‌وجو برای راه حل‌های مبتنی بر هش
گرانی راه حل‌های مبتنی بر SSTable
فقدان ویژگی های ACDI

ویژگی های ACDI

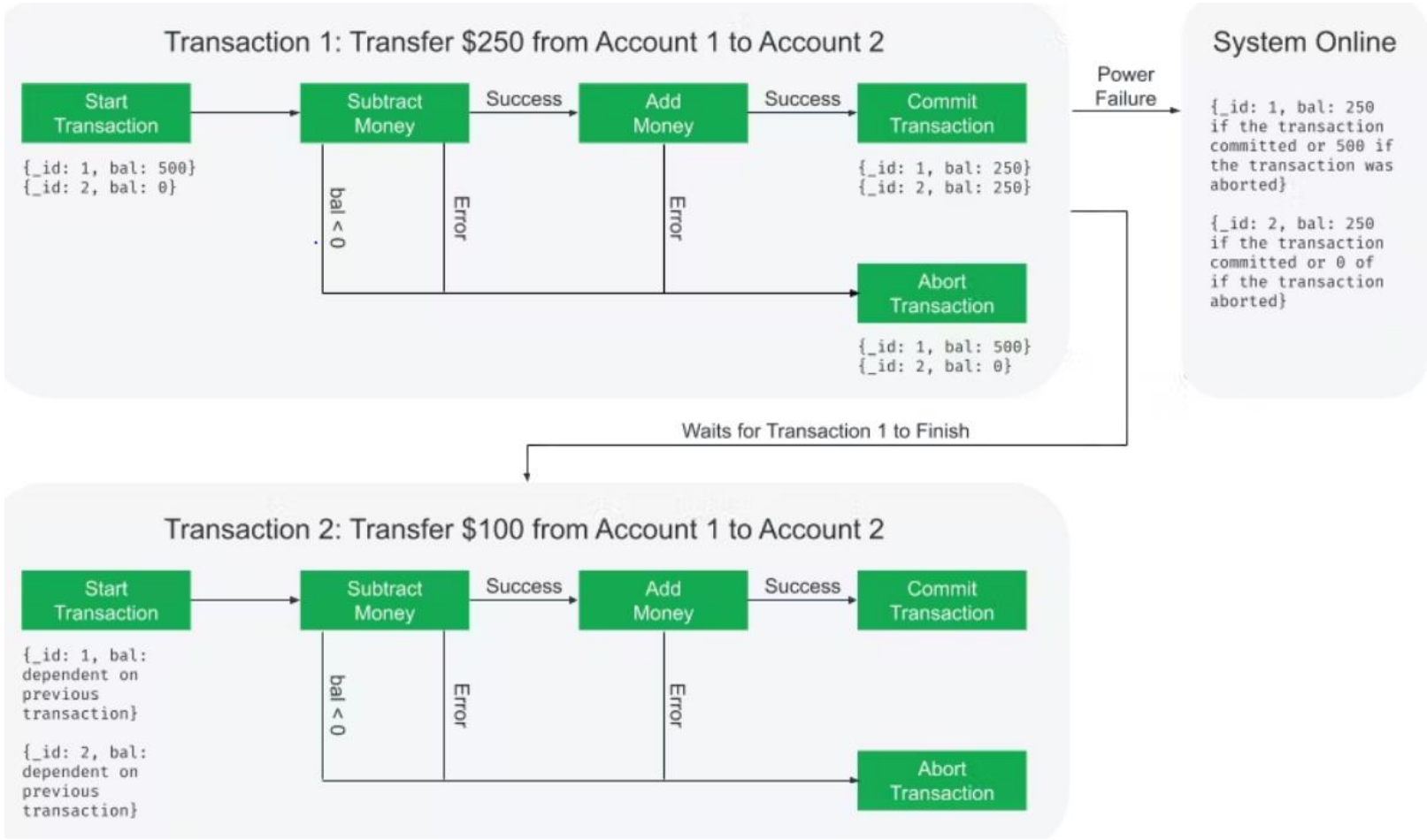
ویژگی‌های ACID در پایگاه داده به مجموعه ای از اصول اشاره دارد که پردازش قابل اعتماد و سازگار تراکنش‌های پایگاه داده را تضمین می‌کند

Atomicity : این ویژگی تضمین می‌کند که یک تراکنش به عنوان یک کار واحد و غیرقابل تقسیم تلقی می‌شود.

Consistency : این ویژگی تضمین می‌کند که یک تراکنش پایگاه داده را از یک وضعیت معتبر به حالت دیگر می‌آورد.

Isolation: این ویژگی تضمین می‌کند که تراکنش‌ها جدا از یکدیگر انجام می‌شوند، حتی اگر به طور همزمان انجام شوند. این از تداخل بین تراکنشها جلوگیری می‌کند.

Durability : این ویژگی تضمین می‌کند که به محض انجام یک تراکنش، تغییرات آن دائمی است و حتی در صورت خرابی سیستم قابل بازگشت نیست.


Transaction 1: Transfer $250 from Account 1 to Account 2
Start Transaction
{_id: 1, bal: 500}
{_id: 2, bal: 0}
Subtract Money
Success
Add Money
Success
Commit Transaction
{_id: 1, bal: 250}
{_id: 2, bal: 250}
bal < 0
Error
Error
Abort Transaction
{_id: 1, bal: 500}
{_id: 2, bal: 0}
Power Failure
System Online
{_id: 1, bal: 250 if the transaction committed or 500 if the transaction was aborted}
{_id: 2, bal: 250 if the transaction committed or 0 of if the transaction aborted}
Waits for Transaction 1 to Finish
Transaction 2: Transfer $100 from Account 1 to Account 2
Start Transaction
{_id: 1, bal: dependent on previous transaction}
{_id: 2, bal: dependent on previous transaction}
Subtract Money
Success
Add Money
Success
Commit Transaction
bal < 0
Error
Error
Abort Transaction

ایجاد خروجی مورد نیاز

مرحله سوم خطوط لوله داده

تولید گزارش‌ها یا فایل‌های خروجی

پایگاه‌های داده SQL عملیاتی سنتی مناسب نیستند زیرا برای بسیاری از پرس وجوهای کوتاه به جای پرس‌وجوهای تحلیلی بزرگ بهینه‌سازی شده‌اند

پایگاه داده‌های NoSQL برای این فرآیند بد هستند، زیرا معمولاً از بیان لازم برای پرس و جو و استخراج داده‌ها پشتیبانی نمی‌کند

چگونه LeanXcale خطوط لوله داده را ساده و تسریع می کند

فراتر از قابلیت‌های معمولی یک پایگاه داده SQL، LeanXcale دارای مجموعه‌ای از ویژگیهای منحصر به فرد است که آن را برای ساده‌سازی و تسریع خطوط لوله داده بهینه می‌کند. این ویژگیها دو اثر دارند:

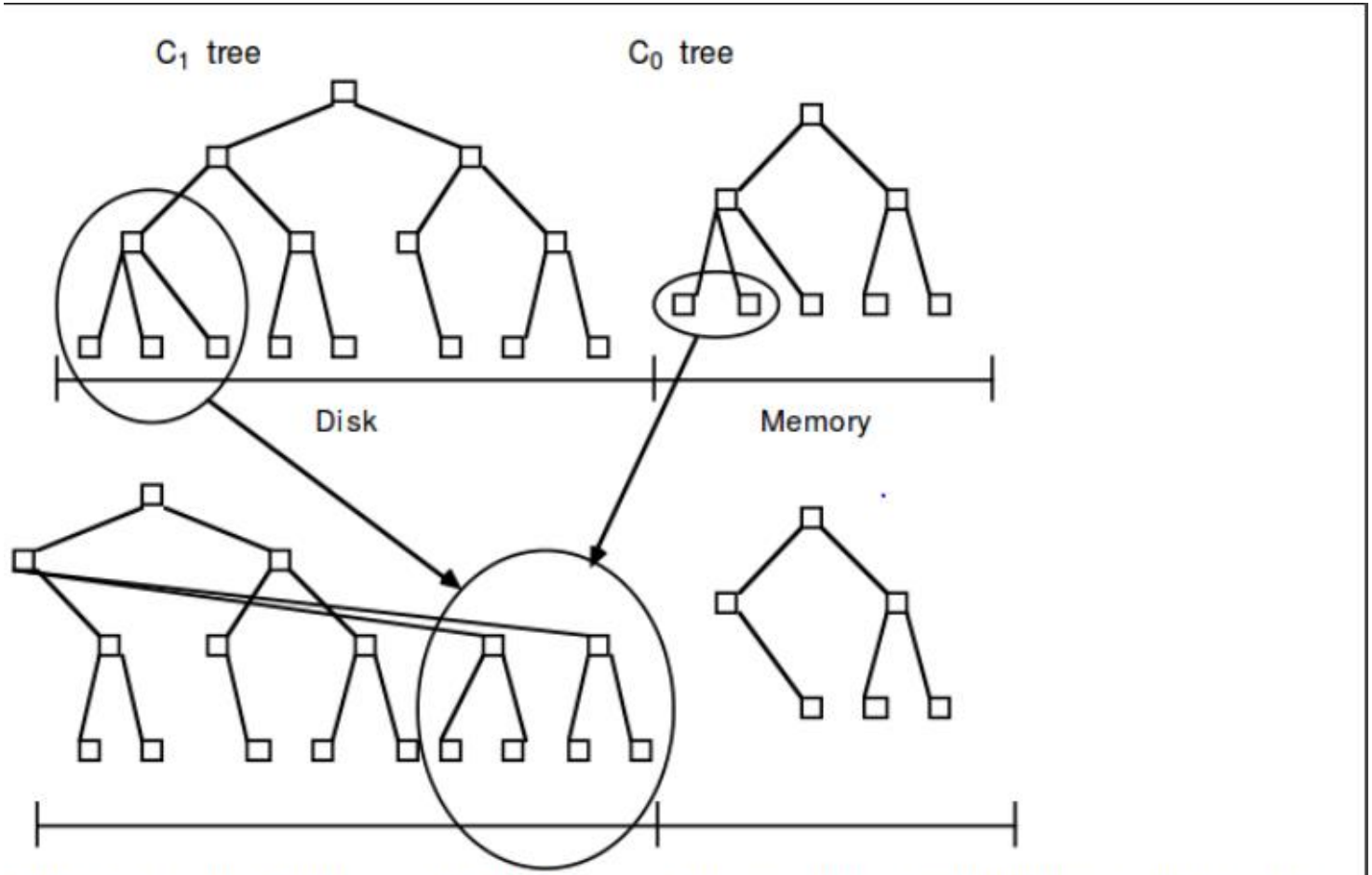
به لطف یک موتور ذخیره کلید-مقدار کارآمد و پارتیشن‌بندی دوبعدی سرعت درج داده‌ها افزایش می‌یابد

قابلیت موازی‌سازی فرآیند در چندین رشته اجرایی و در عین حال اجتناب از قفل شدن پایگاه داده به دلیل تجمع آنلاین و مقیاس‌پذیری افقی خطی

نرخ‌های درج بالا

Leanxcale رویکردی مشابه درختان LSM دارد.

درخت LSM(Merge-Structured-Log) یک ساختار داده‌ای است که در علوم کامپیوتر برای ذخیره و بازیابی موثر داده‌ها استفاده می‌شود. این یک ساختار داده مبتنی بر دیسک است که برای بارهای کاری فشرده، مانند مواردی که در پایگاه داده‌ها و سیستم های فایل یافت می‌شود، بهینه شده است.



2.2. Conceptual picture of rolling merge steps, with result written back to disk

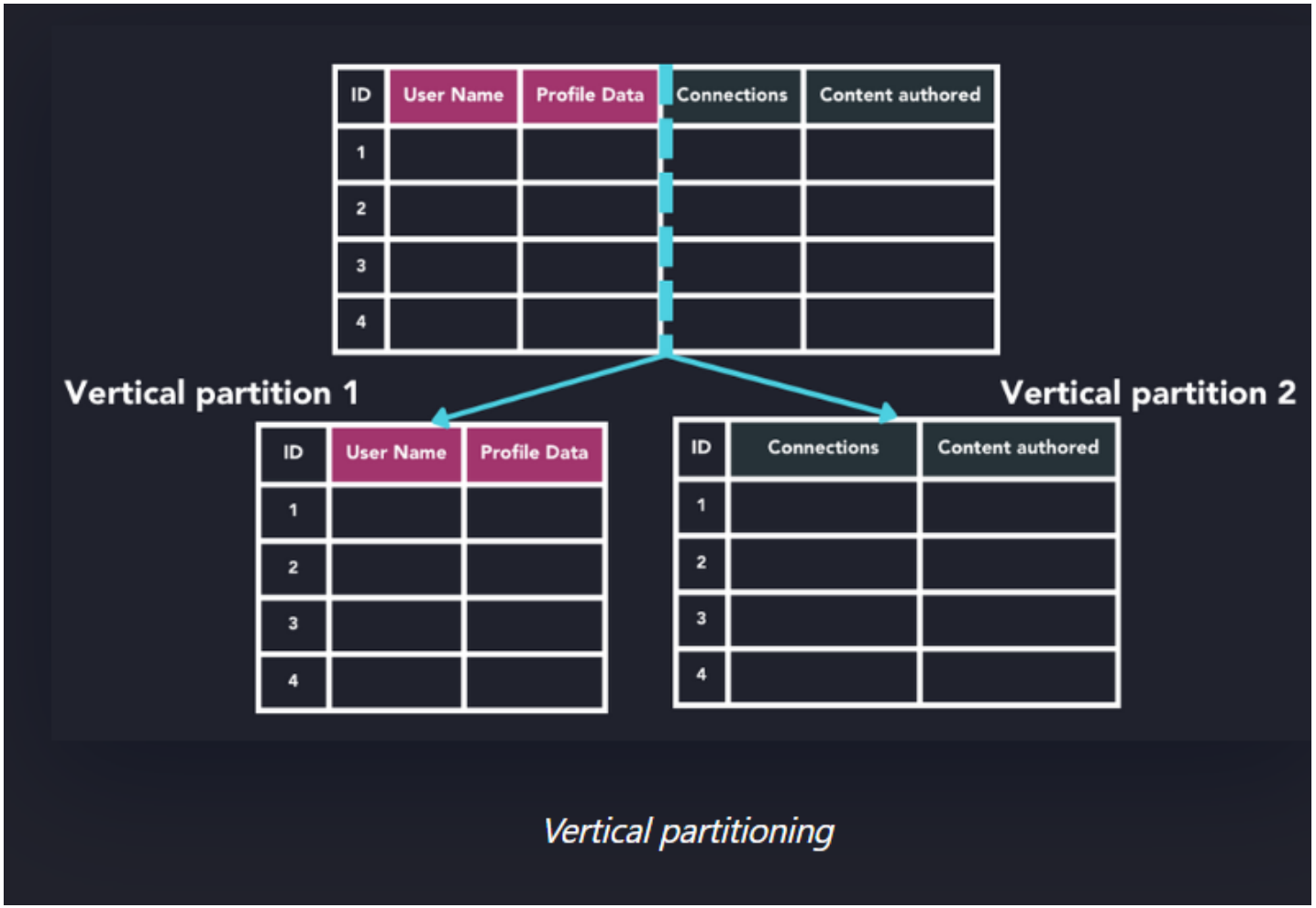
ویژگی‌های منحصر به فرد LeanXcale

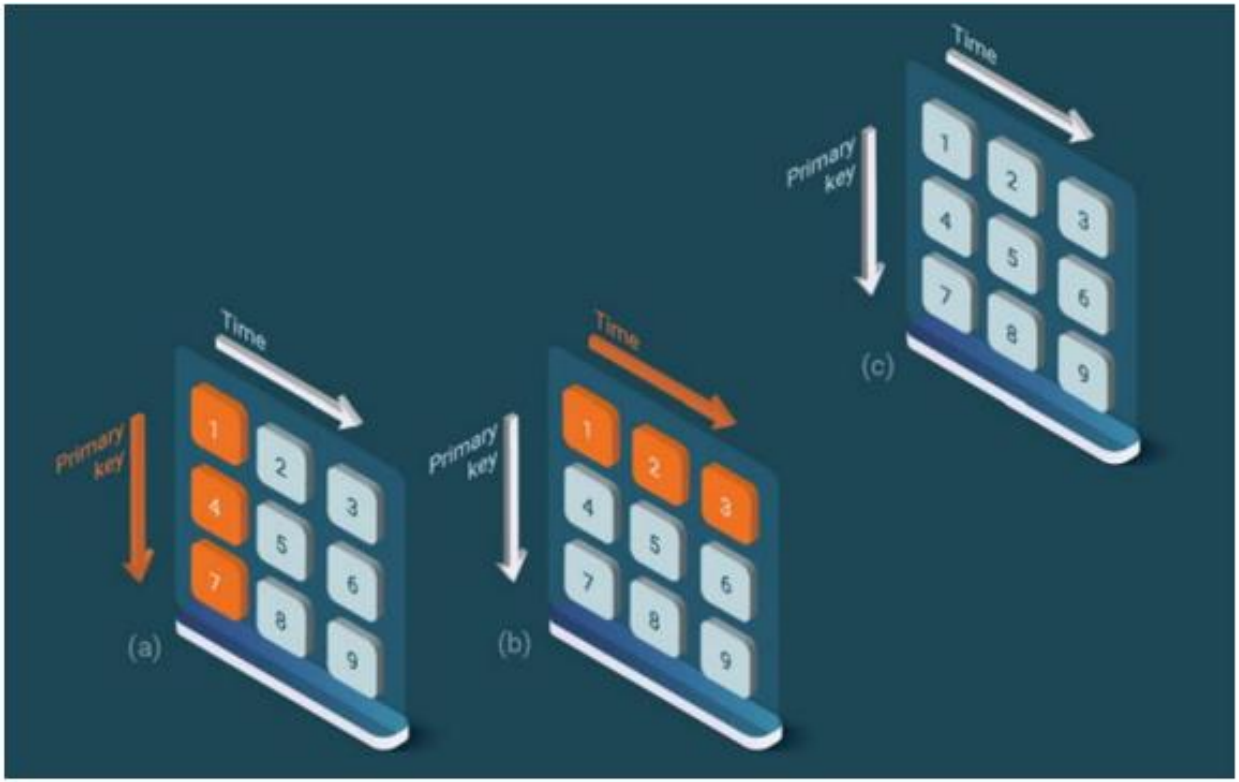
پارتیشن بندی دو بعدی

leanXcale به طور موثر با داده‌های تاریخی طولانی سروکار دارد.

درج داده در جدول اغلب به لحاظ زمانی دارای خصیصه موقتی در اطراف ستون زمان یا ستون خودافزایشگر است .

قطعات داده را به طور خودکار تقسیم می‌کند، بنابراین داده‌ها خیلی بزرگ نمی‌شوند.


ID
User Name
Profile Data
Connections
Content authored
1
2
3
4
Vertical partition 1
Vertical partition 2
ID
User Name
Profile Data
1
2
3
4
ID
Connections
Content authored
1
2
3
4
Vertical partitioning


Time
Primary key
1
2
3
4
5
6
7
8
9
(a)
Time
Primary key
1
2
3
4
5
6
7
8
9
(b)
Time
Primary key
1
2
3
4
5
6
7
8
9
(c)

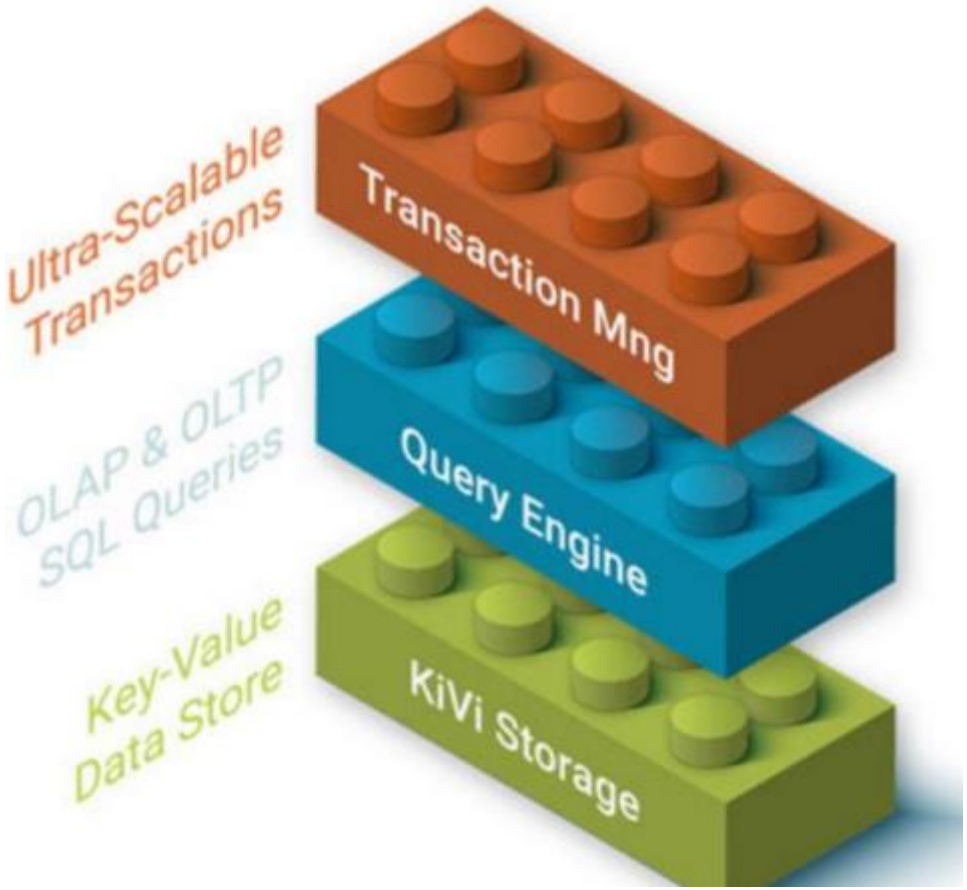
انبوه‌های آنلاین

روش کنترل همزمانی جدید به نام کنترل همزمانی چند نسخه‌ای معنایی

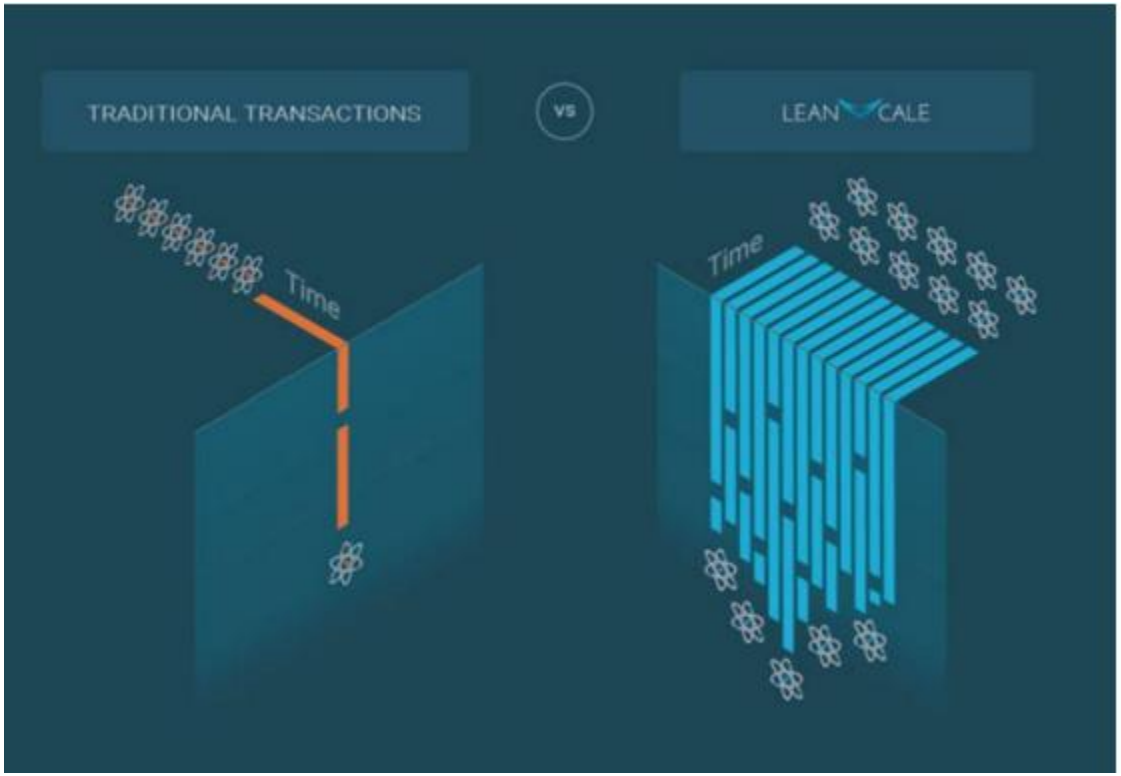
امکان تجمیع در یک ردیف در زمان واقعی با سطوح همزمانی بالا (مثلاً هزاران در ثانیه) بدون هیچ گونه اختلافی و با سازگاری کامل ACID

مقیاس‌پذیری

مقیاس‌پذیری افقی را با مقیاس‌بندی سه لایه پایگاه داده:
موتور ذخیره‌سازی
مدیر تراکنش
موتور پرس‌وجو
لایه ذخیره‌سازی یک فناوری اختصاصی به نام KiVi است که به شکل کلید-مقدار رابطه‌ای است.

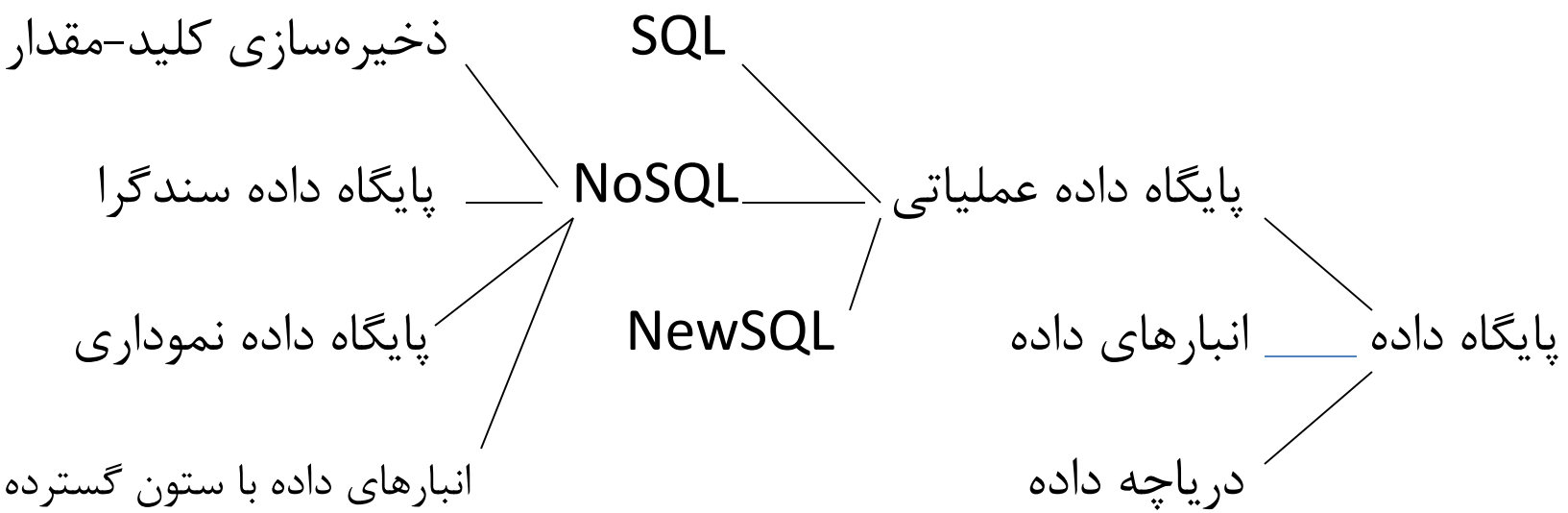

Ultra-Scalable Transactions
Transaction Mng
OLAP & OLTP SQL Queries
Query Engine
Key-Value Data Store
KiVi Storage

اساساً همه تراکنشها را به صورت موازی و بدون هیچ هماهنگی انجام می‌دهد



لوله گذاری

به دلیل استفاده از پایگاه داده‌های مختلف، داده‌ها باید از یک پایگاه داده به پایگاه داده دیگر منتقل شوند، این عمل لوله‌گذاری داده نامیده می‌شود.



پایگاه‌های داده عملیاتی را میتوان بیشتر به سه دسته کلی تقسیم کرد:

پایگاه‌های داده SQL سنتی

محدودیت اصلی مقیاس‌پذیری آنهاست: آنها یا مقیاس‌بندی نمی‌شوند یا فقط به صورت لگاریتمی مقیاس‌بندی می‌شوند، به این معنی که هزینه آنها به طور تصاعدی با مقیاس حجم کاری که قرار است پردازش شود، افزایش می‌یابد.
پایگاه‌های داده NoSQl
میتوانند تضمین‌های سازگاری ACID را فراهم کنند. از سوی دیگر، اکثر آنها می‌توانند مقیاس شوند، اگرچه همه انواع آن این توانایی را ندارند. برخی از آنها می‌توانند مقیاس شوند اما نه به صورت خطی یا نه با تعداد گره‌های زیاد.

NoSQL

ذخیره سازی داده‌های کلید-مقدار

ذخیره سازی داده‌های کلید-مقدار بدون طرحواره هستند
به همین دلیل آنها قابلیت های بسیار کمی برای پرس وجو ارائه می‌دهند

پایگاه‌های داده سندگرا

از داده‌های نیمه ساختاریافته نوشته شده به زبان محبوبی مانند JSON یا XML پشتیبانی می‌کنند. قابلیت اصلی آنها این است که میتوانند داده‌ها را در یکی از این زبانها به طور کارآمد ذخیره کنند و پرس‌وجوهایی را برای این داده‌ها به روشی مؤثر انجام دهند

NoSQL

پایگاه داده‌های نموداری

پایگاه‌های داده گراف در ذخیره‌سازی و پرس‌وجو داده‌های مدل سازی شده به صورت گراف تخصص دارند. داده‌های نموداری که در قالب رابطه‌ای نمایش داده می‌شوند برای پرس‌وجو بسیار گران می‌شوند.

انبارهای داده با ستون گسترده

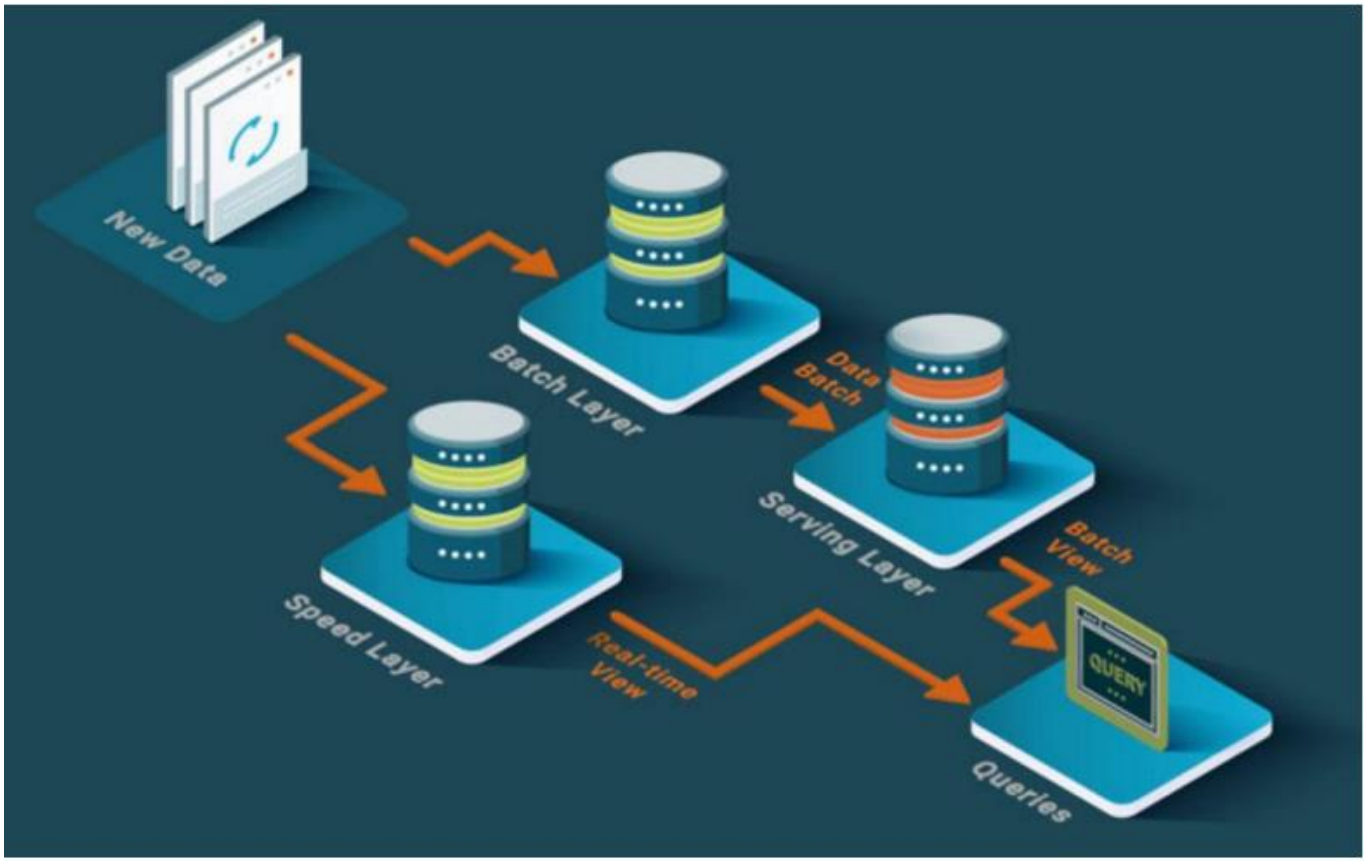
نسبت به ذخیره‌سازی داده‌های کلیدی، قابلیتهای بیشتری را ارائه می‌کنند. آنها معمولاً پارتیشن بندی محدوده را انجام می‌دهند، بنابراین از پرس‌وجوهای محدوده پشتیبانی می‌کنند.

NewSQL

NewSQL یک کلاس از سیستم های مدیریت پایگاه داده رابطه ای است که به دنبال ارائه مقیاس پذیری سیستم های NoSQL برای بارهای کاری پردازش تراکنش آنلاین (OLTP)و در عین حال حفظ ضمانت های ACID یک سیستم پایگاه داده سنتی است.

بسیاری از سیستم‌های سازمانی که داده‌های با مشخصات بالا را مدیریت می‌کنند (به عنوان مثال، سیستم‌های پردازش مالی و سفارش) برای پایگاه‌های اطلاعاتی رابطه‌ای معمولی بسیار بزرگ هستند، اما دارای الزامات تراکنش و سازگاری هستند که برای سیستم‌های NoSQL عملی نیستند تنها گزینه‌هایی که قبلاً برای این سازمان‌ها در دسترس بود، خرید رایانه‌های قوی‌تر یا توسعه میان‌افزار سفارشی بود که درخواست‌ها را روی DBMS معمولی توزیع می‌کرد. هر دو رویکرد دارای هزینه‌های زیرساختی بالا و/یا هزینه‌های توسعه هستند. سیستم‌های NewSQL سعی می‌کنند تضادها را با هم آشتی دهند.

معماری لاندا



لایه دسته‌ای

لایه سرعت

لایه سرویس

فراتر از معماری لاندا

- معماری لاندا به سادگی با پایگاه داده LeanXcale جایگزین می‌شود که تمام قابلیت‌های معماری لاندا را فراهم می‌کند و هیچ گونه پیچیدگی و هزینه توسعه و نگهداری ندارد.

- ناکارآمدی در جذب: همانطور که پایگاه داده رشد می‌کند هر درج نیاز به خواندن یک گره برگ از درخت + B دارد که ابتدا باید یک گره را از حافظه پنهان خارج کرده و روی دیسک بنویسد. این بدان معناست که هر نوشتن حداقل به دو IO نیاز دارد.

- LeanXcale این مشکل را با استفاده از نوع جدیدی از درختان LSM به لطف ترکیب قابلیت‌های SQL و NoSQL با ارائه کارایی ذخیره‌سازی داده‌های کلید-مقدار در مصرف حل می‌کند.

- سهولت در پرس‌وجو : معماری لاندا نیاز به توسعه برنامه‌نویسی هر پرس‌وجو با سه پایه کد مختلف برای هر یک از سه لایه دارد. در LeanXcale، پرس‌وجوها به سادگی در SQL نوشته می‌شوند. پرس‌وجوهای SQL بر خلاف پرس‌وجوهای برنامه‌های در معماری لاندا که نیاز به بهینه‌سازی دستی در سه پایه کد مختلف برای هر یک از لایه‌ها دارند، بهطور خودکار بهینه می‌شوند.

تقسیم داده های تاریخی کنونی

برای حل این مشکل، در LeanXcale، یک الگوی جدید، به نام ذخیره‌سازی داده‌ها در زمان واقعی استفاده می‌شود. این الگو با یک نوآوری که در LeanXcale معرفی می‌شود، حل خواهد شد، یعنی توانایی تقسیم پرس‌وجوهای تحلیلی بر روی LeanXcale و یک انبار داده خارجی. اساساً، قطعات قدیمی‌تر داده را به صورت دوره‌ای در انبار داده کپی می‌کند.

Data marts

مشکل رایج:

اشباع انباره داده

راه حل اول:

ایجاد نماهای کوچک از داده‌های موجود در انبار داده و ذخیره آنها را در پایگاه‌های داده مستقل

Data marts


Data Mart
Data Mart
Data Mart
Data Mart
Data Mart
Data Mart
Data Warehouse

Snapshotting

با Snapshot گرفتن شما می‌توانید وضعیت پایگاه داد ه را در لحظه فعلی ثبت و داخل فایلی نگهداری کنید تا بتوان در زمان مورد نیاز از امکاناتش استفاده نمایید. شاید بتوان گفت به نوعی بکاپ گیری است و می‌توان این فایل را Restore کرد.

یک عکس فوری از صورت دوره‌ای داده‌ها با تازگی یکسان یا بالاتراز داده‌های mart ذخیره می‌شود. مزیت آن این است که کپی به جای اینکه از انبار داده بیاید، مستقیماً از پایگاه داده عملیاتی می‌آید، در نتیجه داده‌های جدیدتری به دست می‌آید.

به لطف کارایی و سرعت دریافت داده‌های LeanXcale، می‌توان عکسبرداری فوری را به جای هفتگی روزانه انجام داد زیرا فرآیندهای بارگذاری که روزها طول میکشید به چند دقیقه کاهش می‌یابد.

جمع‌بندی

LeanXcale، منجر به معماری‌های ساده‌تر با تازگی داده‌ها یا حتی داده‌های بلادرنگ می‌شود. این معماری‌ها به دلیل سادگی، فرآیند توسعه را سرعت می‌بخشند و زمان بین جمع‌آوری نیازمندی‌ها و شروع ایده تا نرم‌افزارهای آماده تولید را کاهش می‌دهند نگهداری آنها مقرون به صرفه تر است زیرا متخصصان کمتر، سرورهای مختلف و مجوزهای پایگاه داده مورد نیاز هستند. علاوه براین، قابلیت‌های پردازشی آن‌ها راه چابک‌تری برای سرعت بخشیدن به فرآیندهای تجاری و در عین حال کاهش ریسک عملیاتی فراهم می‌کند. در نهایت، این معماری‌ها واقعاً می‌توانند از ایجاد جریان‌های درآمدی جدید با ایجاد راه‌های جدید برای برآوردن نیازهای مشتری، با استفاده از نفت امروزی یعنی داده پشتیبانی کنند.

با سپاس از شما