Их хэмжээний өгөгдөл дээр дүн шинжилгээ хийх боломжтой бизнесийн мэдээллийн системийг хэрэгжүүлэх аргачлал

(Дэлгүүрийн бизнесийн мэдээллийн систем)

Амарбаясгалан Цацрал, Намсрай Оюун-Эрдэнэ
Шинэ Монгол Технологийн Дээд Сургууль, Улаанбаатар, Монгол
Хэрэглээний Шинжлэх Ухаан Инженерчлэлийн Сургууль, Монгол Улсын Их Сургууль, Улаанбаатар, Монгол
(a_tsatsral, oyun_erdene79)@yahoo.com

***Хураангуй*—Бизнесийн мэдээллийн систем нь олон тооны эх үүсвэрүүдээс бүтэцлэгдсэн болон бүтэцлэгдээгүй өгөгдлүүдийг нэгтгэн цуглуулж, шинжилгээ хийснээр бизнесийн шийдвэр гаргахад туслах таамаглалуудыг дэвшүүлдэг. Өөрөөр хэлбэл уг системийн тусламжтайгаар их хэмжээний өгөгдлийн цаана нуугдсан олон сонирхолтой мэдээллийг олж илрүүлэх боломжтой юм [1].**

Уламжлалт бизнесийн мэдээллийн системүүд нь зөвхөн тухайн байгууллагын хүрээнд, тодорхой загварын дагуу үүссэн өгөгдөл (бүтэцлэгдсэн) дээр дүн шинжилгээ хийдэг байсан. Гэвч сүүлийн үед шийдвэрлэх өгөгдлийн хэмжээ нь хагас бүтэцлэгдсэн болон бүтэцлэгдээгүй хэлбэртэйгээр эрчимтэй өсөж байна. Тиймээс төрөл бүрийн хэлбэртэй өгөгдлүүдийг нэгтгэн төвлөрүүлж, боловсруулах шаардлага тулгарч байгаа юм [2].

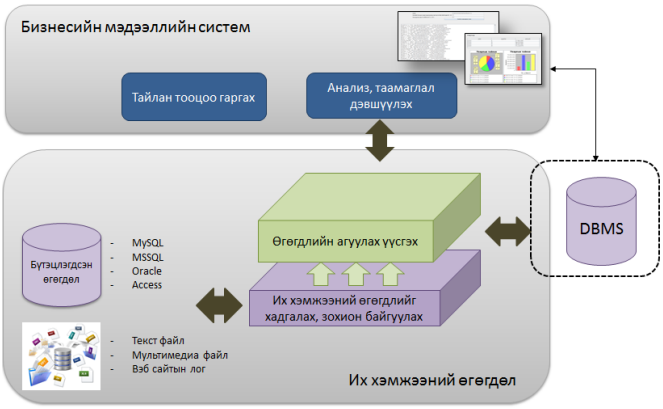
Энэхүү өгүүлэлд их хэмжээний өгөгдлийг боловсруулах боломжтой бизнесийн мэдээллийн системийг хэрхэн хэрэгжүүлэх шийдлийг санал болгож, уг шийдлийг дэлгүүрийн бизнесийн мэдээллийн систем дээр туршсан. Их хэмжээний өгөгдлийг хадгалах, түүнийг боловсруулахад сүүлийн үеийн технологи болох Hadoop болон түүн дээр суурилсан Hive өгөгдлийн агуулах, Sqoop холбоост өгөгдлийн сантай холбох хэрэгсэл, Mahout машин сургалтын алгоритмын санг ашигласан. Эдгээр технологиуд нь бүгд нээлттэй эхийн лицензтэй бөгөөд энгийн үзүүлэлттэй хэдэн мянган компьютер ашиглан өгөгдлийг тархаан байрлуулж, буцаагаад нэг юм шиг ажиллуулж чаддагаараа давуу талтай.

Keywords—өгөгдлийн уурхай; өгөгдлийн агуулах; бүтэцлэгдсэн өгөгдөл; бүтэцлэгдээгүй өгөгдөл; холбоост загвар; хэмжээст загвар;

I. Оршил

1990 оноос хойш дэлхий даяар бүртгэлийн програм хангамж руу анхаарлаа хандуулсны үр дүнд хангалттай их хэмжээний өгөгдлийг цуглуулж чадсан бөгөөд одоо ч өгөгдлийн хэмжээ маш хурдацтайгаар өссөөр байна. Гэвч боловсруулалт хийгдээгүй өгөгдөл хэмжээний хувьд их боловч үнэ цэнэ багатай байдаг. Тиймээс цуглуулсан өгөгдлийг үхмэл байдлаар хадгалахын оронд түүн дээр төрөл бүрийн боловсруулалтыг хийж, ирээдүйг таамаглан, шийдвэр гаргахад ашиглах чиг хандлага бий болж байна.

Энэхүү өгүүлэлд *Зураг* ***1*** –д харуулсан системийн ерөнхий архитектурт тохирох арга технологийн шийдлийг санал болгож, түүнийг дэлгүүрийн шийдвэр гаргахад туслах бизнесийн мэдээллийн системд хэрэгжүүлж, үр дүнг харуулсан.



Зураг 1. Шийдвэрлэх архитектурын ерөнхий зохиомж

Бизнесийн байгууллагууд аливаа шийдвэрийг ирээдүйд хэр зэрэг ашигтай байхыг нь таамагласны үндсэн дээр гаргадаг бөгөөд уг таамаглалыг бизнесийн мэдээллийн системийн тусламжтайгаар гаргах боломжтой. Уг систем нь өгөгдлүүдийг ялгаатай эх үүсвэрүүдээс цуглуулж, зохион байгуулах болон зохион байгуулагдсан өгөгдөл дээрээ шинжилгээ хийж таамаглал дэвшүүлэх гэсэн үндсэн хоёр үйл ажиллагаанаас бүрддэг.

Уг системийг дэлгүүрт нэвтрүүлснээр үйлчлүүлэгчдийн худалдан авалтын ерөнхий зүй тогтол, бүтээгдэхүүнүүдийн хоорондын холбоо хамаарлыг олж илрүүлэх боломжтой. Жишээлбэл ямар бүтээгдэхүүнүүд хоорондоо хамаарал ихтэй байгааг олж тогтоосноор түүнийг хаана юутай хамт байршуулах, цаашлаад юуг юутай хамт сурталчлах, бүтээгдэхүүнүүдийг багцаар нь зарахдаа багцад юу юуг оруулах, ямар төрлийн үйлчлүүлэгчид ямар бүтээгдэхүүнийг санал болгох зэргийг илүү үйлдэл хийж ашиггүй зардал гаргахгүйгээр оновчтойгоор шийдвэрлэх боломжтой болж байгаа юм.

Их хэмжээний өгөгдлийг хадгалах, түүнээсээ өгөгдлийн агуулах үүсгэх, өгөгдлийн агуулахаас таамаглал дэвшүүлэх үйл ажиллагаануудад шаардагдах арга, технологи, алгоритмыг цогцоор нь шийдэж санал болгож байгаа нь тус судалгааны ажлын шинэлэг тал болсон.

II. Судалгаа

A. *Өгөгдлийн сангийн холбоост болон хэмжээст загваруудын харьцуулалт*

Өгөгдлийн сангийн хэмжээст загвар нь холбоост загварын ойлголтуудыг ашигладаг боловч хүснэгт хоорондын холболтын схемээрээ ялгаатай. Холбоост загварын дагуу өгөгдлийн санг үүсгэж байгаа үед нэг хүснэгт хэдэн ч хүснэгттэй холбогдож болдог бол хэмжээст загварын дагуу хүснэгт бүр зөвхөн нэг л төвийн хүснэгттэй холбогддог. Өөрөөр хэлбэл хэмжээст загвараар зохион байгуулагдсан хүснэгтүүдийн холбоо хамаарал харьцангуй бага учраас мэдээлэл гаргаж авахад нэгтгэх үйлдэл бага хийгддэг. Ингэснээр өгөгдлийн сангаас асуух асуулга нь хялбар бөгөөд хурдтай ажилладаг байна.

Хэмжээст загвар нь fact болон dimension гэсэн хоёр төрлийн хүснэгтүүдээс бүрддэг. Fact хүснэгт нь аливаа үйл ажиллагаатай холбоотой бодит тоон мэдээллүүдийг төвлөрүүлэн хадгалдаг бол түүнтэй холбоотой текстэн мэдээллүүд нь dimension төрлийн хүснэгтүүдэд хуваагдан хадгалагддаг [*3*]. Мөн Time гэсэн dimension хүснэгтийг заавал үүсгэх ёстой. Үүнийг цаг хугацаанаас хамаарсан тооцоолол хийхэд ашигладаг. Холбоост загвар болон хэмжээст загваруудын ялгааг ХҮСНЭГТ **1** –д харуулав:

ХҮСНЭГТ 1. Холбоост болон хэмжээст загваруудын харьцуулалт

Өгөгдлийн сангийн систем	**Өгөгдлийн агуулахын систем**
Яг одооны өгөгдлийг хадгалдаг	Түүхэн өгөгдлийг хадгалдаг
Нарийвчилсан өгөгдлийг хадгалдаг	Нарийвчилсан болон нэгтгэн дүгнэсэн өгөгдлүүдийг хадгалдаг
Өгөгдөл нь хувирамтгай	Өгөгдөл нь тогтмол
Мэдээлэл гаргаж авахын тулд дахин дахин боловсруулах үйл ажиллагаа хийдэг	Цаг хугацааны давтамжтайгаар боловсруулах үйл ажиллагаа хийгддэг
Үйл ажиллагаатай холбоотой мэдээллийг хадгалдаг	Судалгаа шинжилгээнд тулгуурласан мэдээллийг хадгалдаг
Тодорхой хэрэглээний програмтай хамтарч ажилладаг	Судалгаа шинжилгээний чиглэлийн програмуудтай хамтарч ажилладаг
Өдөр тутмын шийдвэрийг дэмждэг	Стратеги шийдлийг дэмждэг
Олон тооны хэрэглэгчтэй	Цөөн тооны менежерүүдэд үйлчилгээ үзүүлдэг

Холбоост загвараар өгөгдлийг үр дүнтэй хадгалах боломжтой бол хэмжээст загвараар өгөгдлийг үр дүнтэй авах боломжтой.

B. *Ижил төстэй ажлуудын судалгаа*

Томоохон байгууллагуудыг судалгаа шинжилгээнд зориулагдсан өгөгдлөө хэрхэн, ямар технологи ашиглан хадгалдаг болохыг судалсан.

1) Facebook: Facebook компани анх дүн шинжилгээ хийхэд зориулагдсан өгөгдлүүдийг олон машинуудад тархаан байрлуулахдаа MySQL өгөгдлийн санг ашигладаг байсан бөгөөд Python скриптээр боловсруулалтыг нь хийдэг байжээ. Гэвч тархаан байрлуулсан өгөгдлүүдийг буцаагаад нэгтгэхэд өгөгдлийн санд ачаалал үүсэх, удах, үр дүн муутай байх зэрэг асуудлууд тулгарч эхэлсэн байна. Тиймээс 10 TB хэмжээтэй Oracle өгөгдлийн агуулахыг байгуулсан. Гэвч энэ нь жижиг, дунд байгууллагуудад тохиромжтой шийдэл байсан юм. Учир нь Facebook –д сэтгэгдэл бүртгэгдсэн эхний өдөр л гэхэд 400 GB орчим өгөгдөл үүссэн байна. Тиймээс өгөгдөл цуглуулах болон боловсруулах давхаргаа Hadoop кластераар сольсон байна [*4*]. Ажиллагааны зарчим нь өдөрт цугларч буй хэдэн TB лог буюу бүртгэлийн мэдээллүүдийг цуглуулж HDFS файл системд оруулдаг. Харин HDFS файл систем болон MySQL сервер дээрх өгөгдлүүдийг Hive өгөгдлийн агуулахад нэгтгэх бөгөөд эндээсээ тайлан тооцоо, шинжилгээ хийж, гаргасан үр дүнгүүдээ буцаагаад MySQL болон Oracle серверүүдэд хадгалдаг байна [*5*]. Энэ нь их хэмжээний өгөгдөл дээрх гол тооцооллуудыг Hive өгөгдлийн агуулах дээр хийгээд гаргасан үр дүнг нь олон дахин хандаж авахын тулд энгийн өгөгдлийн сан удирдах систем рүү дамжуулж байгаа гэсэн үг юм. Эцсийн хэрэглэгч бэлэн болсон үр дүнг MySQL болон Oracle серверүүдээс авна.

2) Ebay: Ebay нь 120 сая идэвхтэй хэрэглэгчтэй, өдөрт 300 сая хайлт хийгддэг, 350 сая боломжит бүтээгдэхүүнүүдтэй онлайнаар худалдаа явуулдаг олон улсын байгууллага юм. Тиймээс хэрэглэгчийн дарсан даралт болон бүтээгдэхүүн, гүйлгээ, үйлчлүүлэгч, санал хүсэлт, дуудлага худалдааны өгөгдлүүдийг хадгалж боловсруулалт хийхдээ Hadoop технологийг ашигладаг байна. 2007 онд 4 зангилаанаас бүрдсэн кластер үүсгэж байсан бол 2009 онд 28 зангилаа, 2010 онд 532 зангилаа бүхий кластер үүсгэсэн байна [*6*]. Ажиллагааны зарчим нь Facebook –ийн адилаар Hive өгөгдлийн агуулахад шаардлагатай өгөгдлүүдийг нэгтгэн цуглуулдаг боловч нэг онцлог нь OLAP хэрэгслийг ашиглан өгөгдлийг хялбар аргаар хандаж авдаг байна.

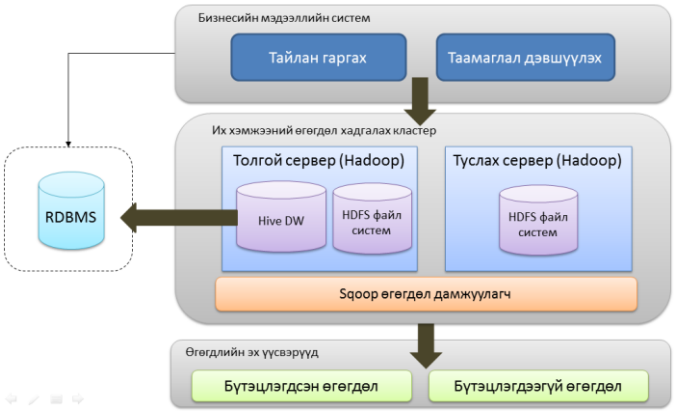
III. Санал Болгож Буй Шийдэл

Их хэмжээний өгөгдлийг хадгалах, зохион байгуулах, түүнийг боловсруулахад шаардагдах арга технологиудыг судалсан бөгөөд Apache –аас гаргасан Hadoop технологийг хэрэгжүүлэхээр сонгосон. Уг технологи нь нээлттэй эхийн лицензтэй бөгөөд бүтэцлэгдсэн, бүтэцлэгдээгүй өгөгдлүүдийн аль алийг нь хадгалж чаддагаараа давуу талтай. Мөн уг технологи дээр суурилсан нэмэлт нээлттэй эх бүхий технологиуд гарч ирсээр байгаа бөгөөд тэдгээр нь уг технологийн ашиглалтыг улам илүү уян хатан болгож өгдгөөрөө давуу талыг бий болгож байгаа юм.

Эхлээд төрөл бүрийн эх үүсвэрүүдээс тайлан болон, таамаглалд ашиглагдах өгөгдлүүдийг цуглуулж HDFS файл системд хадгалах бөгөөд түүнээсээ шаардлагатай өгөгдлүүдийг өгөгдлийн агуулах руу ялгаж оруулна. Өгөгдлийн агуулах нь HiveQL асуулгыг дэмждэгээрээ давуу талтай. Гэвч нэгтгэн дүгнэсэн тайланг шууд Hive өгөгдлийн агуулахаас гаргах нь удаан байдаг. Тиймээс Hadoop болон Hive дээрх өгөгдлүүдийн нэгтгэн дүгнэсэн

үр дүнг MySQL өгөгдлийн санд урьдчилан хуулаад, түүнээсээ тайлан тооцоо гаргавал илүү бага хугацааг зарцуулах боломжтой.

Санал болгож буй архитектурын шийдлийг *Зураг* ***2*** –д харууллаа.



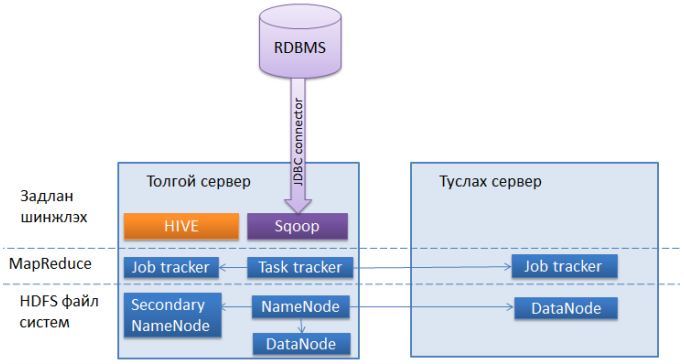
Зураг 2. Хэрэгжүүлсэн архитектурын арга технологийн шийдэл

IV. Дэлгүүрийн Бизнесийн Мэдээллийн Системийн Хэрэгжүүлэлтэд Санал Болгосон Шийдлийг Ашиглах

Санал болгож буй шийдлийг дэлгүүрийн шийдвэр гаргахад туслах бизнесийн мэдээллийн системд хэрэгжүүлж туршсан. Уг шийдлийн дагуу зохион байгуулж хадгалсан өгөгдөл дээр хоёр төрлийн боловсруулалтыг хийсэн. Эхнийх нь Hive өгөгдлийн агуулахаас жил, улирал, сар гэх мэт цаг хугацаанаас хамаарсан багц тайлангуудыг HiveQL асуулгаар гаргаж авсан. Хоёр дахь боловсруулалтад өгөгдлийн уурхайн холбоо хамаарлыг илрүүлэх арга техникийг ашиглан бүтээгдэхүүнүүд хоорондоо хэр зэрэг уялдаа холбоотойгоор зарагдаж байгааг олж илрүүлсэн.

A. Их хэмжээний өгөгдөл хадгалах серверийг бэлдэх

Өгөгдөл хадгалах серверт толгой сервер (master) болон туслах сервер (slave) –ийн үүргээр ажиллах хоёр компьютер ашигласан. Үүсгэсэн Hadoop кластерын ерөнхий бүтцийг *Зураг* ***3***-д харуулав.



Зураг 3. Үүсгэсэн Hadoop кластерын ерөнхий бүтэц

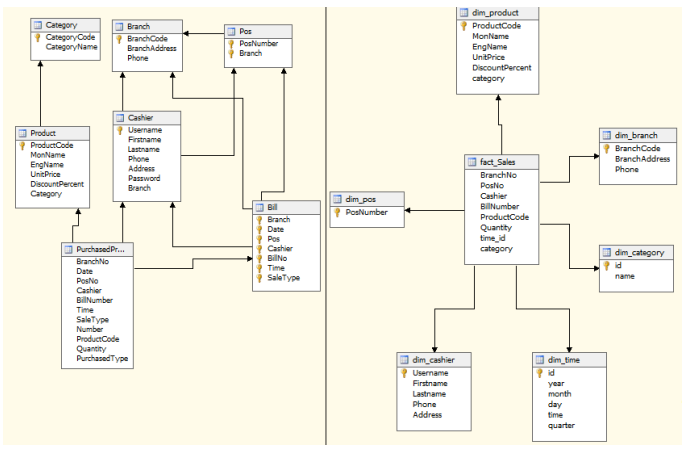
Толгой серверт Hadoop (өгөгдөл хадгалах), Hive (өгөгдлийн агуулах үүсгэх), Sqoop (холбоост өгөгдлийн сантай холбох хэрэгсэл) системүүдийг суулгаж тохиргоог хийсэн бол туслах серверт зөвхөн Hadoop системийг суулгасан.

B. Худалдан авалтуудын мэдээллийг HDFS файл систем рүү оруулж, өгөгдлийн агуулах үүсгэх, тайлан гаргах

Hive дээр өгөгдлийн агуулахыг үүсгэснээр уг агуулахад хадгалагдаж буй бүтэцлэгдсэн болон бүтэцлэгдээгүй өгөгдлүүдийн аль алинаас нь HiveQL асуулга ашиглан хэрэгтэй мэдээллээ гаргаж авах боломжтой болсон. Нөгөө талаас холбоост загвараар хадгалагдсан өгөгдлийг өгөгдлийн агуулах руу оруулахдаа хэмжээст загвар луу хувиргаснаар HiveQL асуулгын бичлэгийг энгийн болгож, асуулгад хариулах хурдыг багасгасан (холбоост болон хэмжээст загваруудын ажиллагааны харьцуулалтыг үр дүн хэсэгт харуулсан).

Өөрөөр хэлбэл Hive дээрх өгөгдлийн агуулахаас цаг хугацаанаас хамаарсан тайлангуудыг гаргахдаа OLAP хэрэгслийг ашиглаагүй боловч түүнд ашиглагддаг өгөгдлийн сангийн хэмжээст загварыг өгөгдлийн агуулахыг байгуулахдаа ашигласан. Маш олон бизнесийн мэдээллийн системүүд цаг хугацаанаас хамаарсан мэдээллүүдийг хялбар бөгөөд хурдан гаргаж авахын тулд өгөгдлийн агуулахыг OLAP хэрэгсэлтэй холбож, түүнээсээ тайлан тооцоог гаргадаг. Гэтэл OLAP хэрэгслүүд нь бүтэцлэгдсэн өгөгдлийн сантай холбогдож (SQL систем), хүснэгтүүдийг нь хэмжээст загварын дагуу хувирган авдаг. Гэвч бидний ашиглаж буй Hive өгөгдлийн агуулах нь холбоост өгөгдлийн сангийн систем биш (NoSQL систем) учраас OLAP хэрэгслийг түүнтэй холбохын тулд нэмэлт арга технологи шаардагдах бөгөөд нөгөө талаас OLAP хэрэгслүүд нь үнэтэй байдаг.

Hive сервер дээр warehouse_star, warehouse_relational гэсэн хоёр өгөгдлийн агуулах үүсгэж, шаардлагатай бүх хүснэгтүүдийг хэмжээст болон загвараар байгуулсан. Үүний үр дүнд бэлэн болсон өгөгдлийн агуулахуудыг *Зураг* ***4***-д харуулав.

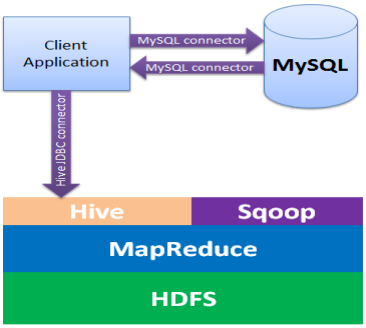


Зураг 4. Дэлгүүрийн өгөгдлийн сангийн холбоост болон хэмжээст загваруудын ялгаа

Hive өгөгдлийн агуулахыг хэмжээст загвараар зохион байгуулснаар доорх давуу талууд бий болсон:

- Энгийн асуулгууд – хэмжээст загварын нэгтгэх (join) үйлдэл нь өндөр нормчилсон холбоост загварын нэгтгэх үйлдлээс илүү энгийн байдаг.
- Хялбаршуулсан бизнесийн тайлан – өндөр нормчилсон холбоост загвартай харьцуулахад цаг хугацааны үечилсэн тайлан гаргахад илүү хялбар.
- Асуулгын гүйцэтгэлийн ашиг - хэмжээст загвар нь зөвхөн унших эрхтэйгээр хандаж, тайлан гаргах боломжтой програмын хувьд гүйцэтгэлийг нь сайжруулдаг. Учир нь уг загвараар үүссэн хүснэгтүүдийн холбоо хамаарал бага байдаг.
- Кубуудыг үүсгэдэг – хэмжээст загварыг бүх OLAP системүүд ашигладаг бөгөөд түүнийг куб үүсгэхдээ ашигладаг.

Өгөгдлийн агуулахаас цаг хугацаанаас хамаарсан тайланг гаргах хэрэгжүүлэлтийг жава програмчлалын хэлээр хийсэн. Hive өгөгдлийн агуулахаас цаг хугацаанаас хамаарсан тайлангийн мэдээллийг авч (HiveQL асуулгыг ашиглах) MySQL сервер лүү хуулах, MySQL серверээс бэлэн болсон тайланг авах (SQL асуулга ашиглах) гэсэн хоёр хэсгээс бүрдэнэ. Тайлан боловсруулах үйл явцыг *Зураг 5*-д харуулав:



Зураг 5. Тайлан боловсруулах үйл явц

C. *HDFS файл системд хадгалагдсан өгөгдлөөс бараа бүтээгдэхүүнүүдийн хоорондын холбоо хамаарлыг илрүүлэх*

Дэлгүүрийн шийдвэр гаргахад туслах бизнесийн мэдээллийн системийн хувьд хэрэглэгчийн худалдан авалтын сагсанд шинжилгээ хийж хамгийн их худалдаалагдаж байгаа бүтээгдэхүүн дундаас хоорондоо хамгийн их хамааралтай бүтээгдэхүүнүүдийг олж тогтоох юм.

Зүйлүүдийн холбоо хамаарлыг илрүүлэх олон алгоритмууд байдаг боловч тэдгээр нь гүйцэтгэлийн хурдаараа ялгаатай байдаг. Их хэмжээний тархаан байрлуулсан өгөгдлөөс холбоо хамаарлыг илрүүлэхдээ Apache –аас хөгжүүлсэн Mahout машин сургалтын алгоритмын санг ашигласан. Уг санд агуулагдаж буй алгоритмууд нь MapReduce зарчмаар боловсруулалтыг хийдэг бөгөөд их хэмжээний тархаан байрлуулсан өгөгдөл дээр ажиллахад зориулагдсанаараа давуу талтай. Холбоо хамаарлыг илрүүлэхэд FPGrowth алгоритмыг ашигласан бөгөөд уг алгоритм нь 2 үе шаттайгаар ажилладаг:

1. FP-Tree үүсгэх
2. FP-Tree –ээс давтагдаж байгаа элементүүдийг шууд авах байдлаар үр дүнг гаргаж авна.

V. Үр Дүн

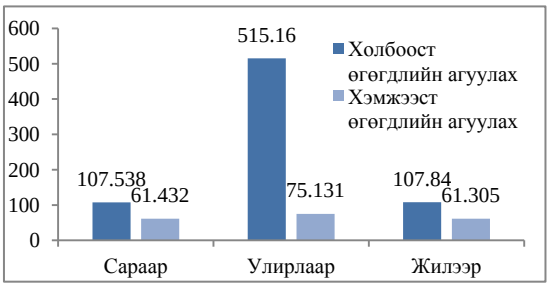
Энэ хэсэгт санал болгосон шийдлийн дагуу хэрэгжүүлсэн дэлгүүрийн бизнесийн мэдээллийн системийн үр дүнг танилцуулсан. *Зураг **6*** –д дэлгүүрийн бизнесийн мэдээллийн системийн хэрэгжүүлэлтийн ерөнхий схемийг харуулав.

Зураг 6. Хэрэгжүүлэлтийн ерөнхий схем

A. *Хэмжээст загвараар зохион байгуулсан Hive өгөгдлийн агуулахын үр дүн*

Үүсгэсэн хэмжээст загвар бүхий өгөгдлийн агуулах нь ***fact_sales, dim_branch, dim_cashier, dim_pos, dim_time, dim_product, dim_category*** гэсэн хүснэгтүүдээс бүрдэх бөгөөд нийт 200057 мөр бичлэг бүхий худалдан авалтын мэдээллийг агуулна. Харин холбоост загвар бүхий өгөгдлийн агуулахын өгөгдөл нь хэмжээст загвар бүхий өгөгдлийн агуулахтай адилхан боловч тэдгээрийн хүснэгт хоорондын холболт нь ялгаатай юм. Уг ялгааг *Зураг **4*** –д харуулсан.

Хэмжээст болон холбоост загвараар зохион байгуулсан өгөгдлийн агуулахуудыг харьцуулахдаа адилхан хариу өгөх асуулгуудыг асууж, хариулах хурд болон асуулгын бичлэгүүдийг хооронд нь харьцуулсан.

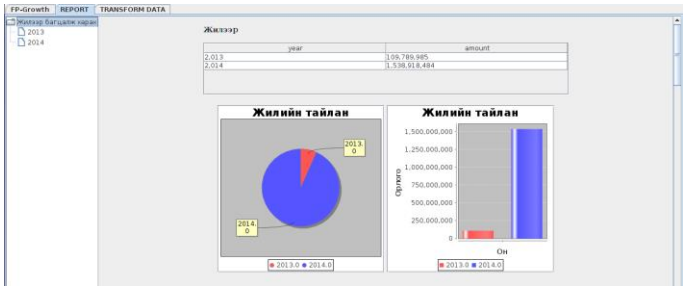


Зураг 7. Хэмжээст болон холбоост өгөгдлийн агуулахуудын харьцуулалт

Асуулгыг сараар, улирлаар, жилээр асуусан бөгөөд сар сараар нь тайлан гаргахад холбоост загвар бүхий өгөгдлийн агуулахаас 107.538 секунд, хэмжээст өгөгдлийн агуулахаас 61.432 секундэд, улирлаар тайлан гаргахад холбоост өгөгдлийн агуулахаас 515.16 секунд, хэмжээст

өгөгдлийн агуулахаас 75.131 секундэд тус тус үр дүнгээ өгсөн. Нийт асуулгын хувьд холбоост загварын асуулга нь хэмжээст загварын асуулгаас харьцангуй их бичлэгтэй болсон. *Зураг 7* –с харахад, нэгтгэн дүгнэсэн үр дүнг гаргахад (зөвхөн унших зорилго бүхий) өгөгдлийн агуулахыг хэмжээст загвараар зохион байгуулсан нь холбоост загвараасаа илүү хурдан гэдэг нь харагдаж байна.

Тайланг MySQL өгөгдлийн санд урьдчилан байрлуулсны дараа тайлан гаргахдаа Hive өгөгдлийн агуулах руу биш MySQL сервер лүү хандана. *Зураг 8*-д жилийн орлогын нийлбэр дүнг харуулсан байна. Мөн үүнтэй адил улирал болон сар тус бүрийн орлогын нийлбэрийг уг серверээс авч харуулж болно. Тайланд хэрэглэгдэх мэдээллийг MySQL сервер лүү урьдчилан боловсруулж хийснээр, тайлан гаргах хурд нэмэгдэж байгаа юм.



Зураг 8. Жилийн орлогыг харуулж буй хэсэг

B. FPGrowth алгоритмын хэрэгжүүлэлтийн үр дүн

Уг алгоритмыг туршихдаа хоёр төрлийн өгөгдлийн эх үүсвэрийг ашигласан. Эхнийх нь уг алгоритмыг туршихад зориулан үүсгэсэн хуурмаг өгөгдөл бөгөөд 304 багана, 1162 мөртэй, хоёр дахь нь бодит өгөгдөл бөгөөд 81 багана 31742 мөрнөөс бүрдсэн.

Хэрэгжүүлсэн FPGrowth алгоритмын үр дүнд бүтээгдэхүүнүүдийн холбоо хамаарал хэр бодитой илэрч байгааг Weka програмын үр дүнтэй харьцуулсан. Weka програмын Apriori болон FPGrowth алгоритмуудын оролтоор дэлгүүрийн шийдвэр гаргахад туслах бизнесийн мэдээллийн системийн өгөгдлийг өгсөн. Эдгээр алгоритмуудыг ажиллуулахдаа minSupport = 0.04 (нийт худалдан авалтын 0.04% –д нь оролцсон байх), confidence = 0.4 (хоорондын хамаарлын хувь нь хамгийн багадаа 40%) байхаар тохируулсан. Уг судалгаагаар хэрэгжүүлсэн алгоритмын үр дүн нь Weka програмын үр дүнтэй нийцтэй гарсан юм.

ХҮСНЭГТ 2. WEKA БОЛОН ХЭРЭГЖҮҮЛСЭН ПРОГРАМУУДЫН ҮР ДҮНГИЙН ХАРЬЦУУЛАЛТ

Холбоотой бүтээгдэхүүнүүд	Weka Apriori	Weka FPGrowth	Hadoop & FPGrowth
HotDogBuns==> HotDogs	0.71	0.71	0.7125
HamburgerBuns==> 98pct.FatFreeHamburger	0.68	0.68	0.6804
Aspirin==> WhiteBread	0.63	0.63	0.6263
Tomatoes==> WhiteBread	0.61	0.61	0.6111
Toothpaste==> WhiteBread	0.6	0.6	0.6018
Toothpaste ==> Eggs	0.56	0.56	0.5648

VI. ДҮГНЭЛТ

Энэхүү судалгааны ажлаар их хэмжээний өгөгдлөөс тайлан болон шийдвэр гаргахад туслах мэдээллийг үр ашигтайгаар гаргаж авах бизнесийн мэдээллийн системийг хэрэгжүүлэх арга технологийн шийдлийг гаргаж, түүнийгээ дэлгүүрийн бизнесийн мэдээллийн системд хэрэгжүүлж туршлаа.

Хэрэгжүүлэлтийг хийхээс өмнө өгөгдлийн агуулахуудын зохион байгуулалтаас шалтгаалж түүний гүйцэтгэл хэрхэн өөрчлөгдөж байгааг туршсан. Туршилтын үр дүнд хэмжээст загвараар өгөгдлийн агуулахыг зохион байгуулснаар, түүнээс асуух асуулга нь холбоост загвараас асуух асуулгыг бодвол энгийн бөгөөд хурдтай ажиллаж чадаж байсан. Тиймээс дэлгүүрийн мэдээллийн системийн өгөгдлийн агуулахыг хэмжээст загварын дагуу үүсгэж тайлан тооцоо гаргах хурдыг нэмэгдүүлсэн.

Дэлгүүрийн шийдвэр гаргахад туслах бизнесийн мэдээллийн системийн хувьд бараа бүтээгдэхүүнүүдийн хоорондын холбоо хамаарлыг илрүүлэх нь маш чухал бөгөөд Apache –аас хөгжүүлсэн Mahout машин сургалтын алгоритмын сангийн FPGrowth алгоритмыг ашигласан. Уг алгоритмын давуу тал нь тархсан машинуудад, зэрэгцээгээр ажиллах боломжтойгоор бичигдсэн учраас Hadoop системтэй илүү зохицож ажилладаг.

Санал болгосон шийдлийн дагуу хэрэгжүүлэлт хийхэд ашиглах програм хангамжууд нь үнэгүй, техник хангамжийн хувьд энгийн үзүүлэлттэй хэдэн мянган компьютер ашиглаж болдгоороо зардал хэмнэх боловч үнэтэй програмуудтай харьцуулахад суулгах үйл явц нь төвөгтэй байдаг.

НОМ ЗҮЙ

[1] Амарбаясгалан Цацрал, Жаргалсайхан Билгүүн, "Тайлан тооцоонд хэрэглэгдэх өгөгдлийг хугацааны хоцрогдолгүйгээр шинэчлэгддэг байхаар зохион байгуулах аргачлал ба жишээ," Хүрэл тоғоот 2014 ЭШХ, Улаанбаатар, 2014.

[2] Okhaide Samson Akhigbe. (April 2014) Business Intelligence - Enabled.

[3] Tsatsral Amarbayasgalan, Bilguun Jargalsaikhan, Otgonnaran O, Oyun-Erdene Namsrai, "Performance Improvement of Mining Techniques: Supermarket's Data," in *FITAT/ISPM 2014*, Chiang Mai, Thailand, 2014, p. 44.

[4] Gordon Brown. (2010, June) blog.redfin.com. [Online]. http://blog.redfin.com/devblog/2010/06/evolving_a_new_analytical_platform_with_hadoop.html#.VI_imivF9-o

[5] Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy, Facebook Data Infrastructure Team Ashish Thusoo, "Hive – A Petabyte Scale Data Warehouse Using Hadoop,".

[6] www.slideshare.net. [Online]. http://www.slideshare.net/madananil/hadoop-at-ebay