هندسة البيانات

Data Engineering

فهد البكري

عندما ترغب شركة أو جهة في الاستفادة من البيانات المتوفرة لديها خاصة إذا كانت كبيرة الحجم (Enterprise) فهم بالغالب يبحثون عن علماء أو محللي بيانات لمساعدتهم. فحينها سيقوم عالم أو محلل البيانات بطلب الحصول على البيانات ليبدأ في تحليل البيانات المتوفرة والمستخرجة من أنظمة وملفات الجهة المستفيدة في تلك الفترة الزمنية وهو يدعو الله أن تكون البيانات صحيحة وكاملة. ثم يبدأ معها مارثون الجري خلف المبرمجين وموظفي الجهة لفهم الأنظمة والبيانات. كما يجد نفسه في صراعات بين من يؤكد صحة البيانات ومن مقتنع أنها خاطئة. ثم تخرج نتائج الرحلة بتحليلات ورسوم بيانية مدهشة. ثم يأتي موظف صغير في قاعة الاجتماعات يقول بأن الأرقام خاطئة!! كيف ذاك والبيانات مسلّمة من قبلكم؟ للأسف جمعت البيانات من قبل شخص غير متخصص، أو أخطأ في استخراج البيانات، أو لم يستخرج جميع البيانات... الخ. كما أن هذه التحليلات والرسوم البيانية تحلل الوضع الحالي والسابق ولا تراعي أن البيانات تنمو وتزداد، فهي غير مرتبطة بمصدر للبيانات (مثل قواعد البيانات) بشكل مباشر. وإذا كانت مرتبطة بمصدر للبيانات فإنها غير مرنة مع تغييرات الأنظمة وإضافة أنظمة جديدة، أو الارتباط بأنظمة جديدة... الخ.

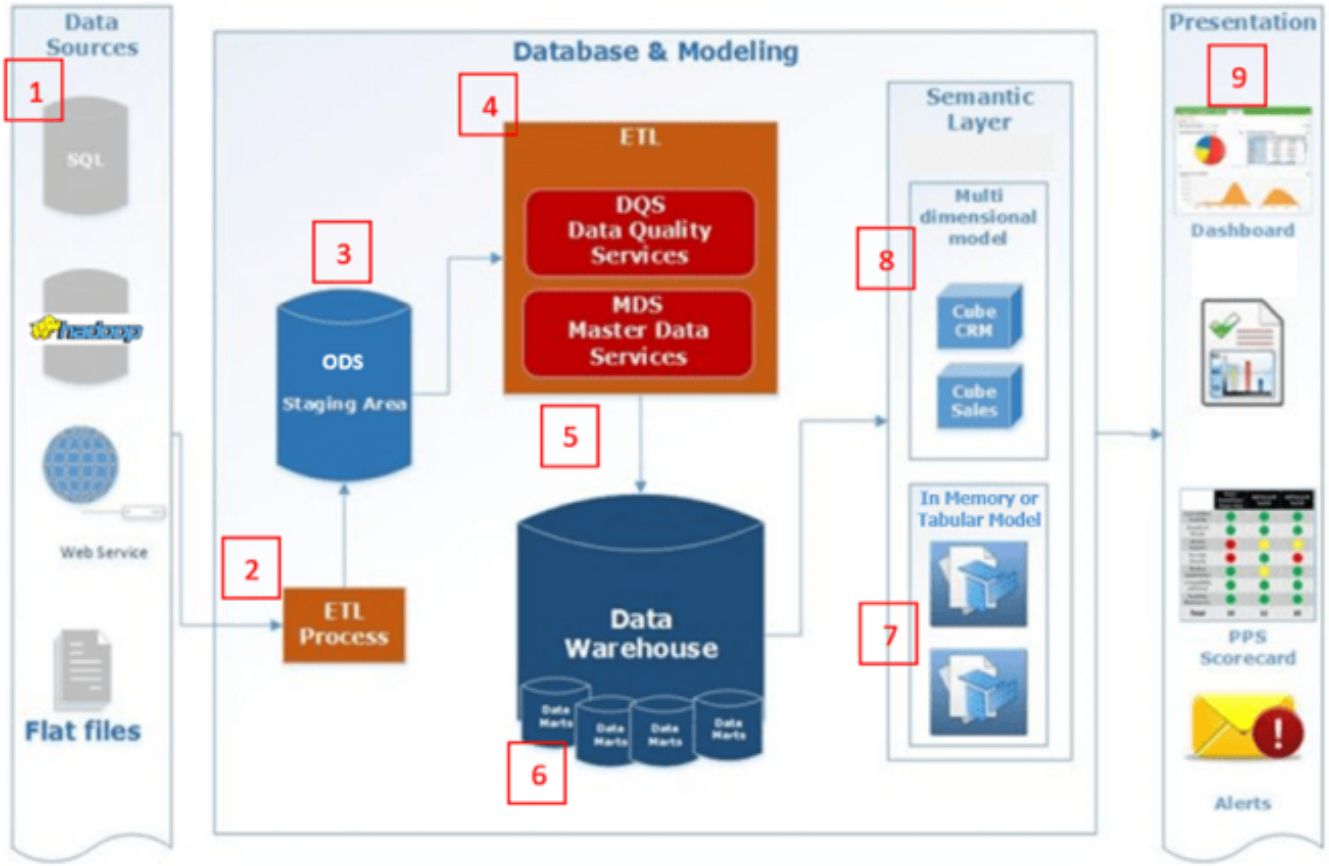
هنا يبرز دور المنقذ؛ مهندس البيانات في دورة حياة البيانات. وإذا أردنا وصف ما تقوم به هندسة البيانات، فهي عملية جمع البيانات من المصدر بشكل مباشر وتنقيحها وتنظيفها وربطها ثم التأكد من جودتها، وربما الربط مع

مصادر أخرى جديدة للتأكد من جودة البيانات، ثم بناء هندسة (هندسة مستودعات البيانات) تسمح بالوصول للبيانات بسرعة ودقة، ثم تجهيزها لمرحلة التحليل. وتراعي هندسة البيانات كفاءة الأداء وأمن المعلومات والخصوصية وحوكمة البيانات. وكذلك التواصل مع كل من له علاقة بالمشروع بشكل تقني كإدارة الخوادم والشبكات وكذلك مصنعي التقنيات المستخدمة لمتابعة كل جديد.

وكإضافة، فهناك برامج تحاول أن تقوم ببعض مهام هندسة البيانات وهي ممتازة منها (Alteryx, KNIME) إلا أن البيانات تكبر مع الزمن لتجد نفسك مجبر للتعامل معها بشكل أكثر احترافية.

ماذا يعمل مهندس البيانات؟

في البداية اعرض لكم مخطط هندسة البيانات مرقم، وسيتم شرح كل مرحلة:



1. مصادر البيانات (Data Sources)

هي الأنظمة وقواعد البيانات والملفات التي سيتم ربطها بنظام البيانات. والمقصود بالربط هو الربط التقني المباشر. وكما هو في المخطط، تظهر المصادر SQL وهي من أهم المصادر وهي تصدر لنا بيانات منظمة في شكل جداول. كما يظهر شعار Hadoop وهو مصدر البيانات غير المنظمة مثل بيانات شبكات التواصل الاجتماعي ومواقع الانترنت وغيرها. وعبارة WEB Service وهي تقنية تساعد على الربط بين الخوادم وتبادل البيانات بينها.

وأخيراً Flat Files والمقصود بها ملفات أكسل وغيرها المنظمة بشكل جداول.

2. استخراج نقل تحميل ETL

هي اختصار Extract Transfer Load، وهي أداة لنقل البيانات تسمح ببرمجة ملفات النقل وإضافة الشروط وتغيير خصائص البيانات وإضافة حقول مركبة والدمج وغيرها. ولها 3 مسميات وظيفية هي:

1. ETL Developer
2. ETL Architect
3. ETL Administrator

يتم بناء ملف برمجي لكل جدول في النظام يسمى Package، فإذا كانت قاعدة البيانات تحتوي 1000 جدول فهذا يعني سوف تنشيء 1000 ملف برمجي، لذا في بداية المشاريع نحتاج عدد كبير من مطوري ETL. يتم تشغيلها كلها بشكل دوري إما يومي أسبوعي أو غيرها. من سيراقب أنها تنجح في كل مرة ولا توجد مشاكل مثلاً في الشبكات أو الخوادم وغيرها؟

إنه بكل تأكيد مدير ETL، يراقب نجاح تشغيل ملفات ETL ويتابع عملية اصلاحها. وأخيراً معماري ETL لا تظهر الحاجة إليه إلا إذا كان فريق ETL كبير، بحيث يدرس الحالات التي لديه ويوزعها على الفريق مع التوجيهات البرمجية اللازمة. وفي الفريق الكبير يصعب شرح كل الأنظمة لمطوري ETL، كما أنهم ليسوا بخبرة واحدة، لذا وجود معماري ETL يكون كالقائد للفريق أمر ضروري.

ولكل ملف ETL مصدر بيانات (Source) ومستقبل بيانات (Destination)، وفي الغالب تكون قواعد بيانات أو ملف اكسل وهي ما تمثله عمليه نقل البيانات.

من خلال المخطط السابق يوجد لدينا 4 مواقع أو خوادم للتخزين البيانات هي:

1. ODS: وهي اختصار لـ Operational Data Source.
2. Data Warehouse: مستودع البيانات.
3. Data Mart: مستودع البيانات المصغر.
4. Semantic/Analysis Server: خادم التحليل.

كل الخوادم السابق ذكرها يتم نقل البيانات بينها عن طريق ETL ماعدا خادم التحليل يرتبط مباشرة إما بالمصدر أو ODS أو DataMart.

3. مصدر البيانات التشغيلي Operational Data Source-ODS

وهي عبارة عن خادم قواعد بيانات يتم أخذ نسخ طبق الأصل من المصدر لعدة أهداف:

- عدم الضغط على مصدر البيانات الفعلي مثل نظام شؤون الموظفين وغيرها.
- تجميع كل المصادر في خادم قاعدة بيانات واحدة يساعد على رواية البيانات بحيث يمكن من عمل Queries مرتبطة من أكثر من مصدر ومقارنة البيانات.
- توحيد التقنية يساهم في تحسين الأداء.
- يمكن استخراج تقارير وبيانات بشكل مستعجل منها، ما يسمى بـ AD hoc Reports.
- يتم تجربة كثير من الاختبارات التقنية والهندسية من خلال هذه الخادم دون التأثير على مصادر البيانات.

ODS هو أول مرحلة في تجميع البيانات، ويجب أن تكون خام مثل المصدر تماماً. ثم تأتي بعدها عدة مراحل كتنظيف البيانات وتجهيز مستودعات البيانات. وبكل أسف كثير من الجهات في مجتمعنا تسميها مستودع بيانات وهذا خطأ كبير فهي فقط مرحلة أولية من تجميع البيانات.

4. تنظيف وجودة البيانات

في هذه المرحلة يتم تنظيف البيانات وهي عبارة عن تصحيح الأخطاء الإملائية أو تغير الرموز الى مسميات أو استبدال NULL بمصطلح "غير معرف" وحذف الزائد من الرموز البرمجية وتوحيد اشكال التواريخ وغيرها.

أما جودة البيانات فهي التأكد من دقة البيانات عن طريق ربطها بمصدر آخر أو عرضها على خبير أو مقارنتها مع تقارير أخرى، مثل عندما يكون في قاعدة بيانات جامعة ما أن عدداً من الطلاب أعمارهم 3 سنوات، فهذه بيانات خاطئة لا يمكن تصحيحها إلا عن طريق جمعها مرة أخرى من الطلاب أو التواصل مع جهة حكومية موثوقة أو الرجوع للملفات الورقية. وفي هذا مجال واسع لدرجة أنه يوجد من هو متخصص في هذا المجال بمسمى وظيفي Data Quality.

في المخطط MDM هو اختصار Master Data Management وهو عبارة عن تحديد مصدر المعلومة الأكثر موثوقية. مثال، لديك عدة أنظمة فيها كلها جدول للجنسيات أيها سوف يكون المصدر الموحد للمعلومة؟ خاصة إذا كان ترميز الجنسيات مختلف. وربما تضطر إلى بناء واحد جديد.

تتم هذه المرحلة عن طريق ETL. ولكن في المرة السابقة التي نقلت ETL البيانات من المصدر إلى ODS كانت عبارة عن كود برمجي عادي، بالعادة تكون مهمة المبتدئين، أما في مرحلة التنظيف والجودة تكون معقدة وتحتاج من لدية خبرة كبيرة بهندسة قواعد البيانات.

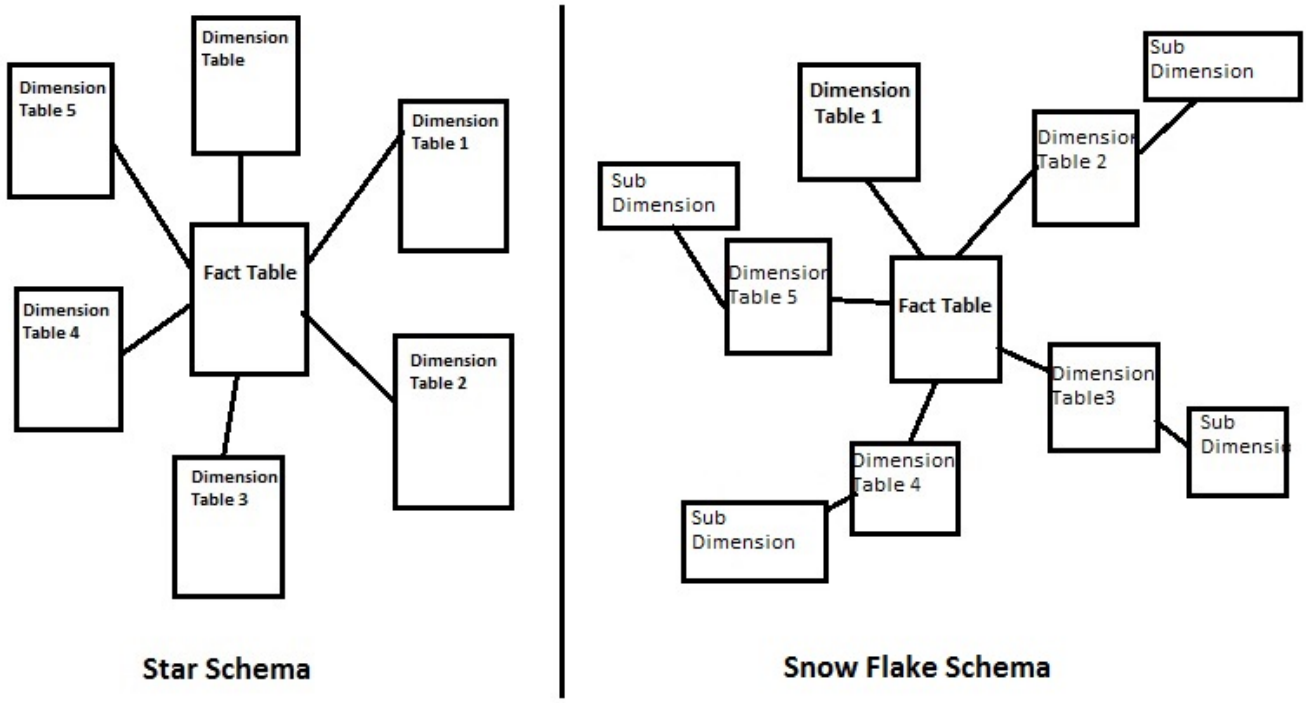
وأخيراً في هذه المرحلة، نبني ملفات ETL جديدة (Package) غير التي تم بنائها في ODS فتجد عدد ملفات ETL يصبح كبير ويحتاج الى إدارة ومتابعة تشغيلها اليومي، لذا هي عملية شاقة وأي خلل فيها سيؤثر بشكل

مباشر على البيانات.

5. مستودعات البيانات Data Warehouses

لا يسعني شرحها بشكل مبسط، هي عالم كامل جميل وفيه ذكاء كبير في التعامل مع البيانات. تبنى على أساسين: **الأول** الحقيقة (Fact) (ويظهر في برامج تصميم لوحات المعلومات باسم Measure) **والثاني** البعد (Dimension). ولتبسيط المسألة، كل حقل في قاعدة البيانات يمكن تجميعه يعتبر Fact، وكل حقل نص أو تاريخ يصف هذا الرقم يعتبر Dimension. مثل عدد الموظفين ومعدل الأعمار ومجموع الرواتب كلها Facts، أما إذا تم استعراضها حسب الجنس مثلاً أو المنطقة أو المؤهل العلمي فتعتبر Dimensions.

تبنى على أساس جداول تتكون من Facts وأخرى Dimensions، وأشهر مدرستين في هندسة قواعد البيانات هما مخطط النجمة Star Schema ومخطط خلية الثلج Snow Flake Schema. لا يسع المجال لشرحهما هنا، وسيتم الحديث عنها في مقالة منفصلة.



يعمل في هذا المجال كل من معماري قواعد البيانات Data Warehouse Architect ومصمم نماذج البيانات Data Modeler. والمعماري هو القائد لمجموعة من مصممي نماذج البيانات.

6. مستودعات البيانات المصغرة Data Mart

هي باختصار تجميع جزء معين من مستودعات البيانات الرئيسية، يتم بناؤها لعدة أهداف رئيسية:

1. مستوع البيانات الرئيسي يجمع كل البيانات، وبالعادة لا يتم عرض كل البيانات في تقرير أو لوحة معلومات واحدة. لذا يتم تجهيز Data Mart تجمع فيها البيانات التي تكون هدف لعرضها في التقرير وهذا بهدف تحسين الأداء.
2. مستخدم التقارير ولوحات المعلومات لا يحق له الاطلاع على كل ما في مستودع البيانات، لذا من باب الخصوصية تبنى له Data Mart خاصة تلبي احتياجاته.
3. لا يمكن وضع كل الشروط والطلبات على مستودع البيانات، فمن الذكاء الهندسي أن يتم تقسيم البيانات في مجموعات على شكل Data Mart.
4. عند طلب العميل إضافة بيانات خاصة له أو يرغب في إضافة بعض المعادلات فإنه يتم بناء Data Mart له لينفذ عليها ما يشاء دون المساس بمستودع البيانات الرئيسي.

لذا وجود Data Mart مفيد جداً هندسياً وأداءً وخصوصيةً وأمناً، وكذلك مرونةً في التفاعل مع طلبات العميل.

واخيراً، يجب أن تعلم أن نقل البيانات من مستودع البيانات إلى Data Mart يتم أيضاً عن طريق ETL.

7. تحليل البيانات المجدول Tabular

هذه التقنية متوفرة في عدة أنظمة بمسميات مختلفة أشهرها Tabular و In Memory وغيرها. هذه التقنية تدعم بناء التقارير من قاعدة البيانات بشكلها الطبيعي المسمى Normalization، وهو مخطط قاعدة بيانات أي نظام مثل نظام شؤون الموظفين والمالية. وأغلب تطبيقات تصميم لوحات المعلومات مثل Tableau وPower BI يكون مدمج فيها هذه التقنية. ومن خلالها تتمكن من الربط بين الجداول أو تغير المسميات وأنوع الحقول والدمج بين الحقول، وغيرها الكثير.

نحن مهندسي البيانات نستخدم هذه التقنية في حالات محدودة مثل التقارير المستعجلة، أو إذا كان النظام الذي نحن بصدد تحليله صغير لا يستحق بناء مستودع بيانات له. هذه التقنية سريعة في الأداء ولكن تستخدم الذاكرة (RAM) بشكل كبير، لذا لا ينبغي الإفراط في استخدامها. ولكن للأسف توجد جهات حكومية تعتمد على هذه التقنية بشكل كامل.

8. تحليل البيانات متعدد الابعاد Multi-Dimensional

هذه التقنية هي إبداع بكل معنى الكلمة، وتعرف بعدة أسماء أخرى مثل مكعب البيانات (Cube) وكذلك OLAP اختصار Online Analytical Processing، هذه التقنية تعتمد على أساس الحقيقة Fact والبعد Dimension بما يعرف بـ Denormalization. حيث تجمع كل Fact ثم تربطه مع كل الابعاد وبكل أشكالها، وتجهيزها للرد على أي سؤال ممكن ومحتمل. مثال، لديك قاعدة بيانات مدرسة بكل بيانات الطلاب، هذه التقنية تجهز إجابات الأسئلة الممكن استخراجها بشكل تلقائي. فإذا أردت معرفة الطلاب المتأخرين

في دراستهم حسب الوضع الاجتماعي لهم مع ربط البيانات مع أداء معلميهم، فهذا السؤال هو معد مسبقاً من قبل هذه التقنية. والمبرمج فقط جهز البيانات لتتصل بهذه التقنية لتعمل هذا الجهد الرائع. يطلق على من يعمل على هذه التقنية BI Developer.

9. بناء التقارير ولوحات المعلومات والمؤشرات Report & Dashboard

في البداية، وجدت اختلاف في ترجمة Dashboard فمنهم من يطلق عليها لوحة المعلومات ومنهم من يسميها لوحة المؤشرات. وإذا تمعنت في المعنى فإن أي معلومة في تقرير هي في الأساس مؤشر لشيء ما، لذا أعتقد أن الإسمين مقبولين. في هذه المرحلة يكون الدور متبادل بين مهندس البيانات وعالم ومحلل البيانات، فالكل يسهم في بناء التقارير. إلا أن في الجهات المحترفة تجد متخصص في واجهة المستخدم UI يساعد على تصميم التقارير لتظهر بشكل جذاب.

الخاتمة

هندسة البيانات يعمل فيها أكثر من شخص، كل منهم يعتبر مهندس بيانات. وهي بالمسميات التالية دمج فيما بينها حسب حجم المنشأة:

1. ETL Developer
2. ETL Architect
3. ETL Administrator
4. Data Modeler
5. Data Warehouse Architect
6. BI Developer
7. Dashboard Designer

وفي بيئتنا الحالية تجد من يجيد أكثر من تخصص. ويوجد عجز في كل المهارات السابقة حالياً من حيث المهتمين بالعمل كمهندس بيانات. وللحق وجدت مهندسي بيانات لديهم مهارات عالم بيانات. لكن لم أجد عالم بيانات يستطيع العمل مهندس بيانات خاصة في القطاعات الكبيرة.

مهندس البيانات هو من يأخذ على عاتقه المهمة الصعبة ويفسح المجال لعالم البيانات أن يبدع. مهندس البيانات لا يستطيع العمل من غير عالم بيانات وكذلك عالم بيانات لا يستغني عن مهندس البيانات. وإذا فقد أحدهما يحدث خلل كبير وكثير من المشاريع تفشل.

في الأخير أدعوا مجتمع الحاسب الآلي للاهتمام في هذا المجال والتخصص فيه.