**به نام خدا**

**عنوان موضوع : لیست مشکلات انباره داده فناوران و حوزه هوش تجاری**

**تهیه کننده : حمید شمس اشراق (کارشناس ارشد مدیریت فناوری اطلاعات)**

**لیست مشکلات :**

**1-شفاف سازی در SSIS(ETL) و SSAS(OLAP) به منظور ارائه تحلیلی از معماری و زیر ساخت و نوع طراحی و پیاده سازی انباره داده که در حال حاضر مستندات آن از طرف فناوران موجود نیست.**

**ابتدا باید کلیه منابع اطلاعاتی که در اختیار فناوران بوده و به آنها دسترسی دارد اعم از سرورهای بانک های اطلاعاتی و کلیه بانک های اطلاعاتی و کلیه جداول و فیلد ها و ارتباطات بین جداول و به طور کلی Data Model های مربوطه که Data WareHouse از آنها استفاده می کند مستند سازی شده و توضیح داده شوند . همچنین پروژه یا پروژه های مبتنی بر بسته های SSIS که به صورت Job ها یی روی Sql Server اجرا می گردند بطور کامل مستند سازی و آموزش داده شوند .فرآیندهای SSIS باید گویا و قابل فهم باشد.**

**همچنین در صورت استفاده از Cube ها ، پروژه یا پروژه های مبتنی بر Cube مستند سازی شده و کاملا تشریح گردد. همچنین آموزش های لازم در زمینه MOLAP های بکار رفته در Cube ها و شاخص های کلیدی عملکرد نیز انجام پذیرد.**

**نحوه اجرای گزارش گیری سلف سرویس از طریق Analysis Service بطور کامل مستند سازی و تشریح گردد.**

**همچنین مفاهیم SSIS و SSAS بطور کامل به تیم مربوطه آموزش داده شود که این امر می تواند در حین توضیحات و تشریح بسته های SSIS و Cube های داده صورت پذیرد.**

**البته در صورتیکه از Cube ها استفاده نشده است مفاهیم SSIS و بسته های ان که یه صورت Job روی SqlServer تعریف شده اند کاملا تشریح گردد.**

**در مورد SSIS در ابتدا مفاهیم رویه های Extract و Transform و Load توضیح داده شده سپس نحوه ساخت بسته های SSIS و پیمایش آن در محیط توسعه مانند Visual Studio همچنین اجرا و استقرار بسته های SSIS کاملا توضیح داده شود.**

**در مورد SSAS در ابتدا پرس و جو های چند بعدی در T-SQL با استفاده از توابع تحلیلی همچنین مقدمه ای بر ساخت مکعب ها در Analysis Service و روش های فراهم سازی گزارش گیری سلف سرویس با اجزای متفاوت Sql Server تشریح شود.**

**در مورد انباره داده باید مستندات معماری زیر ساخت و نوع طراحی و پیاده سازی انباره داده تشریح گردد.**

**2-وجود تعدد جداول بالا در مارت های مختلف به عنوان مثال هم اکنون حدود 873 جدول در انباره داده فناوران موجود می باشد که تحلیل جداول و فیلد ها و ارتباطات بسیار زمان گیر بوده و طراحی Query ها را با پیچیدگی مواجه می کند .با توجه به تعدد زیاد جداول ، برای بدست آوردن یک گزارش یا داشبود ، نیاز به Join ها و SubQuery های زیاد می باشد که این امر در کندی اجرای Query و همچنین کندی در اجرای گزارش یا داشبورد در ابزار Powerbi را بدنبال خواهد داشت.(این قضیه در شرکت تست شده است).**

**چرا باید جداول زیادی در Data WareHouse در مارت های مختلف داشته باشیم با توجه به اینکه یکی از اهداف طراحی و پیاده سازی جداول Data WareHouse نرمال سازی نکردن داده ها و تکراری بودن بخشی از داده ها به خاطر درگیر نشدن در Join های بین جداول می باشد به نظر می رسد حجم تعداد جداول می تواند کاهش یافته و سرعت اجرایQuery ها را افزایش دهد.**

**به عنوان مثال می توان در بعضی از بخش های طراحی انباره داده به جای طراحی دانه برفی از طراحی ستاره ای استفاده کرد تا حجم جداول ابعاد کاهش یابد.یا به جای گذاشتن کلید جداول ابعاد در جداول واقعیات از شرح کلید استفاده کرد تا هنگام طراحی Query ها با ارجاع به شرح فیلد نیاز به Join اضافی نباشد تا هم Query ساده تر گردد و هم سرعت اجرای Query بالارفته و هم طراح با پیچیدگی Query مواجه نباشد.**

**3-نبود بعضی از فیلد ها به صورت سرجمع در داخل یک جدول به عنوان مثال با توجه به اینکه یک داشبورد ترکیبی از عملکرد آمار مدیریتی صدور و خسارت داریم ، می توان در جدول آمار صدور فیلدهای سرجمع خسارت همان رکورد موجود باشد و نیاز به Join نباشد که هم اکنون در بعضی از مارت های انباره داده فناوران موجود نمی باشد.**

**به عنوان مثال در مارت اتومبیل بخش ثالث و بدنه این مسئله رعایت شده است و در جداول صدور ثالث و بدنه سرجمع مبالغ خسارت موجود بوده است که خیلی کمک به محاسبه ضریب خسارت و اجتناب از محاسبات پیچیده و**

**Join های زیاد در جداول انباره داده می کند.همچنین سرعت اجرای Query را بالا می برد.**

**چرا در بعضی از مارت ها مانند اتومبیل این قضیه رعایت شده است و در بعضی از مارت ها متاسفانه موجود نیست.**

**4-بعضی از جداول دارای رکوردهای بالایی می باشند مثلا جدول FactAmarBN حدود 72 میلیون رکورد و جدول FactHV دارای 16 میلیون رکورد می باشد که طبعا سرعت Query را کاهش می دهد آیا توجیهی برای این حجم اطلاعات وجود دارد؟آیا نمی توان حجم اطلاعات را کاهش داد؟**

**چه مکانیسم ها یا روش هایی برای کاهش اطلاعات بعضی از جداول مانند جداول فوق وجود دارد ؟**

**با توجه به اینکه حجم بالای داده ها باعث کندی بارگذاری روی سرور شده ، از طرفی سرعت اجرای Query ها را نیز کاهش می دهد .اگر چه بهینه نوشتن Query ها می تواند تا حد زیادی از کندی بکاهد ولی در حجم داد های بالا مانند اعداد ذکر شده باز هم کندی به دنبال دارد .**

**به عنوان مثال انتخاب فقط یک فیلد از از یک جدول فوق ، ساعت ها به طول می انجامد در صورتیکه حتی یک Join برای این جدول وجود ندارد.طبعا Join این جدول با جداول دیگر زمان بیشتری خواهد برد که کار طراح را با کندی روبرو می کند.**

**5-نبود عملیاتی برای بهبود عملکرد انباره داده ، از قبیل Indexing و Partitioning و Clustring و ...**

**بر روی کلیه جداول انباره داده که خود در بالا بردن سرعت اجرای Query ها نقش بسزایی دارد.به عنوان مثال زمانی که فقط روی جدول FactAmarBN جستجو می کنیم بیش از 30 دقیقه طول می کشد تا اطلاعات واکشی گردد.چرا عملیات فوق بر روی جداول انباره داده صورت نگرفته است؟**

**با توجه به اینکه انباره داده ها ممکن است در دوره های زمانی مختلف بازنگری یا دارای تغییرات در سطح طراحی باشند ممکن است نیاز مجدد به عملیات فوق داشته باشند . عملیات فوق نه تنها در سطح فیلدهای اطلاعاتی بلکه در سطح پایگاه داده ها نیز باید صورت پذیرد مثلا Shrink کردن پایگاه داده به صورت Job روی**

**Sql Server و هر نوع عملیاتی که برای بهینه سازی پایگاه داده مناسب می باشد.**

**6-استفاده نکردن از OLAP و CUBE های اطلاعاتی به جای OLTP که خود می تواند در سرعت بالای گزارشات تاثیر گذار باشد.البته فناوران طی جلسه ای که داشتیم ، توجیهاتی بابت این امر داشتند مثلا تاخیر در پر شدن اطلاعات در CUBE ها علاوه بر تاخیر پر شدن جداول طی فرآیند ETL حدود یک شبانه روز که خود این امر ، OnLine بودن اطلاعات را با مشکل مواجه می کند.**

**همچنین طراحی انباره داده باید کامل و بدن تغییر باشد تا به سراغ Cube برویم اگر نه بارها باید پروژه Cube را دچار تغییرات قرار دهیم چون همه چیز سمت پروژه Cube انجام می گیرد که خود یک چالش است .**

**مثلا اگر ابزار PowerBi را به Cube وصل کنیم همه عملیات را باید سمت پروژه Cube انجام دهیم به شرط اینکه نخواهیم تغییراتی در انباره داده بدهیم در غیر اینصورت پروژه Cube باید دوباره دچار تغییرات گردد که بسیار زمانگیر بوده و زمان طراحی گزارش را بالا می برد زیرا ایتدا زمانی نیاز هست تا توسط ETL داده های انباره داده پر گردد و سپس زمانی هم برای پر شدن Cube که اصطلاحا Cube Process می گویند باید سپری گردد.**

**7-یکی از مباحث مهم تست عملکرد انباره داده می باشد که در زیر نمونه ای را مثال می زنم.**

**فاز تست شامل (تست انبار داده که شامل بهره وری انبار داده و فرآیندهای دسترسی و صحت داده ها و مدل بین منبع اصلی و OLAP و تست ETLها شامل صحت انبار داده پاکسازی شده و نشده و بهره وری روالهای ETL و Performance Tuning در مبدا و مقصد و تست گزارشات و تست یکپارچه سیستم شامل تست فشار روی انبار داده با حداکثر کاربران و تست فشار روی سرور گزارشات و تست کل سرورها و سیستم ها در حالت Full Load ) .**

**تمامی تست های فوق باید توسط فناوران روی انباره داده صورت پذیرد تا از کارایی انباره داده مطمئن شویم.**

**8-نام گذاری جداول و فیلدها در بعضی جاها کاملا English و در بعضی جاها کاملا Finglish و در بعضی جاها ترکیبی از English و Finglish طراحی شده اند و یک الگو و فرمت دهی منظم برای نام گذاری رعایت نشده است.با توجه به تعداد بالای جداول و تعداد بالای فیلدهای هر یک از جداول بهتر است یک الگو را همه جا رعایت کنند که در حال حاضر به این شکل نیست.**

**همچنین طول کاراکتر های نام بعضی از جداول و فیلدها طولانی می باشد که بهتر است کوتاه و مختصر باشد تا حجم کاراکتر های تولید شده در Query را زیاد نکند.**

**همچنین نام فیلد های کلید خارجی از یک الگو پیروی کند که در حال حاضر این طور نیست .**

**به عنوان مثال در بعضی جاها در انتهای نام فیلد از کلمه Id استفاده شده و در بعضی جاها نشده است .این امر باعث می شود طراح با جستجو در ارتباطات جدول رابطه کلیدهای اصلی و خارجی را کشف کند که بسیار زمانگیر می باشد.**

**9-سه مسئله مهم در مدیریت داده های هوش تجاری عبارتند از :**

**1- صحیح بودن داده ها**

**2-بروز بودن داده ها**

**3-آنلاین بودن داده ها از طریق وب سایت،موبایل،تبلت،تلویزیون هوشمند و....**

**که موارد فوق باید از سمت فناوران راستی آزمایی گردد.**

**بعضا اطلاعات با 2 روز یا 3 روز تاخیر بروز می شوند که مسئله online بودن را خدشه دار می کند.**

**با توجه به اینکه برای رئیس هیئت مدیره و مدیر عامل و مدیران ارشد سازمان 3 مسئله فوق بسیار حیاتی بوده و می تواند روی تصمیم گیری های نهایی مدیران تاثیر گذار باشد.زمان ETL یا زمان Cube Process باید طوری باشد که حداکثر اطلاعات با یک روز تاخیر بروز شوند.**

**10-به نظر اینجانب برای هر داشبود مورد نیاز بیمه آسیا می توان فقط یک جدول وجود داشته باشد بطوریکه Dimension ها و Fact ها همگی داخل آن باشد و اگر چنین جدولی موجود نیست ، فناوران آن را برای ما ایجاد کند تا احتیاج به Join ها و SubQuery های متعدد نباشد .**

**ممکن است تکرار بخشی از اطلاعات داشته باشیم که این امر در انباره داده به خاطر ویژگی Denormalize مجاز می باشد. مثلا اگر حدود 50 داشبورد مهم مورد نیاز باشد می توان فقط با 50 جدول سر و کار داشت که همین امر پیچیدگی نوشتن Query ها را پایین برده و سرعت اجرای گزارش در ابزار را بالا می برد.**

**البته این مورد پیشنهاد حقیر بوده و در صورت عدم توافق در مورد فوق باید جداول درگیر برای هر داشبورد یا گزارش مربوطه به حداقل برسد و حتی الامکان همه فیلدهای مورد نیاز ابعاد در جدول واقعیت به صورت متنی باشد تا احتیاج به کمترین Join باشد که در حال حاضر در بعضی موارد این امر رعایت نشده و در بعضی موارد رعایت شده است.**

**11-نیاز به یک مشاور یا مدرس که در صورت داشتن اشکالات یا ابهامات چه در زمینه طراحی Query ها و چه در زمینه ابزار PowerBI بتوان به او ارجاع کرد و در کمترین زمان بهترین پاسخ را بدست آورد تا تیم BI به تدریج تجربه کافی را کسب کند.همچنین فناوران پاسخگوی نیازهای ما باشد به عنوان مثال فیلدهای طول و عرض جغرافیایی که باید در جدول cmn.DimLocation اضافه می شد تا کنون انجام نشده است یا جدول FactLocationPortfoEstimate دارای داده نمی باشد و .....**

**وضعیت فعلی انباره داده فناوران :**

**در حال حاضر حدود 873 جدول در انباره داده فناوران موجود می باشد که ترکیبی از مارت های جدید و مارت های قدیم رشته های بیمه ای بوده است.**

**مارت های جدید با نام Schema برابری می کند که اول هر نام اسکیما یک کاراکتر n به معنی New بکار رفته است.به عنوان مثال ml به معنی مارت مالی قدیم و nml به معنی مارت مالی جدید می باشد.**

**در حال حاضر با توجه به آموزش مارت های مختلف توسط شرکت کارانه گر از سوی فناوران ، مارت های زیر تدریس شده اند که DataModel هر کدام از آنها در اختیار شرکت بیمه آسیا قرار گرفته است.**

**1-مارت مشترکات و مدیریتی(cmn و amr) آموزش دیده توسط آقایان (اشراق و سمیعی و عاشورزاده)**

**2-مارت مالی (nml) آموزش دیده توسط آقایان (سمیعی و عاشورزاده)**

**3-مارت اتومبیل شامل ثالث و بدنه(ncr) آموزش دیده توسط آقایان (اشراق و عاشورزاده)**

**در حال حاضر مارت مالی فقط شامل طراحی جداول و ارتباطات می باشد ولی دارای داده نمی باشد که بارها پیگیری کردیم ولی متاسفانه تا کنون داده های آن توسط بسته های SSIS پر نشده است.**

**همچنین مارت اتومبیل شامل ثالث و بدنه آموزش داده شده است ولی کلا این مارت در انباره داده موجود نمی باشد نه طراحی آن و نه داده آن که باز هم بارها به فناوران متذکر شده ایم و تا کنون اقدامی صورت نگرفته است.**

**با توجه به توضیحات فوق بقیه مارت های رشته های بیمه ای مانند درمان و .... باید از سوی فناوران طراحی و پیاده سازی گردد تا بعد از آموزش در ابزار Power BI پیاده سازی گردد.**

**در حال حاضر با توجه به آموزش های دیده شده و مشکلات بیان شده فوق در مارت های مالی و اتومبیل در ابزار PowerBI گزارشات و داشبوردهایی با توجه به مارت های مشترکات و مدیریتی تهیه شده است که عبارتند از :**

**1-گزارش کلی صدور و خسارت**

**2-گزارش وضعیت تسویه بیمه نامه های ثالث**

**3-گزارش خسارات مستقیم و غیر مستقیم**

**4-گزارش ضریب خسارت**

**5-گزارش مقایسه صدور و خسارت بر اساس سال**

**6-گزارش مقایسه صدور و خسارت بر اساس صدور بیمه نامه-سالانه و ماهانه و روزانه**

**7-گزارش پیش بینی پرتفو**

**وضعیت آتی انباره داده فناوران :**

**به نظر اینجانب با توجه به مشکلات و چالش های ذکر شده می توان فناوران را در جریان قرار داد تا اصلاحاتی در انباره داده انجام دهند البته به شرطی که در مارت های آموزش دیده تاثیری نگذارد و در مارت های باقی مانده این موارد را در صورت توافق انجام دهند.**

**در ضمن مارت های درمان و عمر و پس انداز و عمر و حوادث و آتش سوزی و مسئولیت و مهندسی و باربری باید توسط فناوران در انباره داده طراحی و پیاده سازی شده و به تیم BI از سمت بیمه آسیا آموزش داده شود.**

**همچنین با توجه به KPI های طراحی شده توسط مدیریت ریسک و بهبود فرآیند ها ، انباره داده فناورارن باید جوابگوی نیازهای بیمه آسیا باشد تا بر اساس آن داشبورد هوشمند مدیریتی طراحی گردد.**

**موفق باشید**