[SLIDES 1 & 2]

Après avoir brièvement présenté la démarche d'élaboration du dashboard, ces notes s'attachent à présenter le contexte de l'étude au travers de l'audience visée et des objectifs, avant de décrire comment les données ont été sourcées et organisées pour l'étude. On présentera ensuite les différentes vues sur le dashboard produit à partir de ces données, avant de conclure sur les prochaines étapes la stratégie.

[SLIDE 3]

Tout d'abord, donc, la démarche d'élaboration du dashboard. Elle a consisté en 5 étapes, partant de l'identification des besoins du donneur d'ordres et de leur formalisation dans un blueprint, et se poursuivant dans la définition d'objectifs et d'ébauches de visuels au travers d'un mock-up pour validation par notre donneur d'ordres. Nous avons ensuite sourcé les données pertinentes pour l'analyse et les avons organisées dans une base de données MySQL après avoir géré les différentes anomalies en utilisant Python dans un notebook Jupyter. Nous avons enfin construit les différents visuels dans Power BI et avons analysé et interprété les résultats.

[SLIDE 4]

Dans le contexte de cette étude, nous nous adressons à un bailleur de fonds auquel nous devons démontrer notre expertise dans 3 domaines, à savoir la création de services d'accès à l'eau potable, la modernisation de services existants et l'activité de consulting auprès de gouvernements sur les politiques d'accès à l'eau. La mise en valeur de ces 3 domaines nous a amenés à considérer un certain nombre d'indicateurs pertinents repartis sur 3 niveaux de granularité : mondial, régional et national. Nous nous sommes assurés que ces 3 niveaux de granularité étaient représentés pour chacun des 3 domaines d'expertise, comme le montrent les extraits du blueprint sur la droite de cette diapositive. Par exemple, pour le domaine de la création de services, on a considéré l'évolution dans le temps de la répartition de la population et le taux d'urbanisation mondial, régional et national, l'urbanisation étant un des facteurs déterminant non seulement le besoin de services d'accès à l'eau potable mais aussi d'assainissement. Dans le domaine de la modernisation des services, qui est notre 2ème domaine d'expertise, on a considéré par exemple l'accès de la population à l'eau potable (environ 85% au niveau mondial) et aux services d'assainissement (environ

25% au niveau mondial), avec une granularité urbaine/rurale là où les données existaient, et ce encore une fois sur les 3 vues au niveau mondial, régional et national.

[SLIDE 5]

Enfin pour le 3ème domaine, celui du consulting, on a considéré des indicateurs tels l'évolution de la stabilité politique ou la situation du pays par rapport à la moyenne mondiale des populations ayant accès à l'eau potable et aux services d'assainissement. On a enfin considéré d'autres indicateurs communs à nos 3 domaines d'expertise, comme le filtrage des pays ayant une stabilité politique inférieure à -2 (moins 2) ce qui, compte-tenu de la méthodologie de calcul de cet indice, revient à dire que nous avons exclu de notre étude les 2.3% de pays les plus instables. En effet, les services d'accès à l'eau potable et d'assainissement requérant des investissements très importants en termes d'infrastructure, ils nécessitent une volonté politique et une capacité d'exécution dans le temps que les pays en situation de forte instabilité ne présentent généralement pas.

[SLIDE 6]

Ces différents indicateurs ont donc été mis en place dans un mock-up afin de les mettre en perspective sur les différentes vues et de prendre en compte le besoin d'interactivité des utilisateurs (ici, notre audience qui est un bailleur de fonds). Un certain nombre de filtres modifiables par l'utilisateur ou de sliders ont donc été inclus, sur le choix de la région ou du pays par exemple, ou sur la période historique considérée. Nous aurons l'occasion de démontrer ces fonctionnalités dans la 3ème partie lorsque nous aborderons l'étude des résultats. Pour chacune des vues mondiale, régionale et nationale, nous avons inclus des cartes, permettant un repérage visuel immédiat pour l'utilisateur. Pour chaque indicateur, le graphique approprié a en outre été choisi afin de présenter les données de la façon la plus claire et la plus lisible possible, par exemple des courbes d'évolution historique pour certaines données, des gauges pour la comparaison des pays à la moyenne mondiale ou encore des graphiques en barres pour la répartition des populations et des nuages de points pour les analyses bivariées.

[SLIDE 7]

Passons à présent au sourcing et à l'organisation des données. Les données originales étaient présentées sous forme de fichiers Excel extraits des sites internet

de la FAO, de l'OMS et de la Banque Mondiale. Les données contenues dans ces fichiers présentaient un certain nombre d'anomalies : granularité pas disponible de façon cohérente pour tous les pays, informations manquantes pour certaines années (comme l'indice de stabilité politique qui est absent pour tous les pays en 2001 par exemple : il n'a juste pas été publié) et fichiers présentés avec des données partiellement pivotées avec les catégories figurant en ligne et non en colonne, ce qui rend l'identification des clés primaires problématique, mais qui présente aussi le désavantage de multiplier les valeurs manquantes quand les niveaux de granularité sont différents pour certaines variables. Il a donc fallu retraiter ces anomalies en vue de les éliminer ou au moins de les minimiser.

[SLIDE 8]

Afin d'organiser les données, nous avons choisi d'utiliser une base de données MySQL, dont le schéma est présenté à droite de cette diapositive. Une fois le schéma créé dans MySQL Workbench, il a été fait usage de la fonction « Forward Engineer » pour générer le code de création de la base ainsi que la base elle-même et les différentes tables, le tout en conformité avec la 3NF. Les tables de la base ne contiennent donc aucune donnée calculée, et les totaux et pourcentages ont été retirés au profit de données exprimées en nombres entiers de personnes.

[SLIDE 9]

Voici un aperçu de la base finale telle que créée dans MySQL Workbench, avec dans les encadrés à gauche les différentes tables et les colonnes qui représentent nos différentes variables.

[SLIDE 10]

Afin d'alimenter la base de données, les fichiers Excel initiaux ont été uploadés, inspectés et nettoyés en utilisant des dataframes Python dans un notebook Jupyter. En particulier, on a vérifié l'absence de doublons ainsi que l'unicité des clés primaires en vue de jointures entre les différentes tables, la présence de valeurs NULL a été minimisée en dé-pivotant certaines tables, les champs calculés (pourcentages par exemples) ont été «dé-calculés » afin de respecter la 3NF, les données de population ont été présentées en individus plutôt qu'en milliers d'individus, on a vérifié le type casting des données et renommé les colonnes en conformité avec les intitulés dans la

base pour faciliter l'import et enfin on a retiré les colonnes de total là où une granularité était présente, avant de joindre les données des 5 fichiers initiaux retraités en 3 tables, prêtes à alimenter la base de données dans MySQL. L'alimentation à proprement parler a été effectuée avec la librairie MySQL Connector de Python après avoir exporté les données nettoyées au format .csv. On peut voir sur la droite de cette diapositive le début du code Python avec notamment les différentes librairies utilisées et les instructions de connexion à la base MySQL.

[SLIDE 11]

L'élaboration du dashboard a nécessité l'utilisation d'un outil spécialisé. Ces logiciels ayant des fonctionnalités, des prix et une facilité d'utilisation parfois très différents, notre choix ici a été guidé par la popularité de ces logiciels parmi les grandes entreprises françaises, 2 logiciels se détachant clairement des autres : Tableau et Power BI – nous avons choisi ce dernier, entre autres, en raison de la puissance de l'architecture Microsoft qui le supporte.

L'outil Power BI a donc été connecté directement à la base de données créée dans MySQL : on voit sur la droite de cette diapositive un extrait du menu Power BI permettant cette connexion. Les données ayant été « dé-calculées » pour mise en conformité avec la 3NF, on a réintroduit dans Power BI, là où c'était nécessaire, des colonnes calculées, et on a également ajouté les relations entre les clés primaires des tables là où elles n'étaient pas détectées automatiquement par le logiciel. On a par exemple également regroupé l'indice de stabilité politique des pays en bins pour améliorer la lisibilité des cartes, et ajouté des champs calculés de somme ou de pourcentage.

[SLIDE 12]

La première vue que nous avons produite sur le dashboard est une vue mondiale synthétique, qui présente les données de croissance de la population mondiale sur la période 2000-2017 ainsi que l'évolution de l'accès à l'eau potable et aux services d'assainissement sur la même période au travers de plusieurs graphiques, ainsi que plusieurs indicateurs synthétiques qui concernent l'année 2016 exclusivement, qui est l'année la plus récente pour laquelle nous avons un jeu de données complet : en particulier, nous avons choisi de représenter la population totale ainsi que l'accès total à l'assainissement et à l'eau potable, ce qui nous permettra ensuite comme on le verra

dans la vue pays de comparer la situation nationale de chaque pays à la moyenne mondiale. La stabilité politique a quant à elle été représentée sur une carte après avoir été groupée en bins pour plus de lisibilité au niveau du code de couleurs, qui donne une visualisation immédiate des zones de conflit et d'instabilité marquée (en noir sur la carte). Sur cette première vue, on a donné la possibilité à l'utilisateur d'interagir avec les données au travers d'un slider qui permet de zoomer sur le graphique en barres que l'on voit en bas à gauche de la vue, pour mettre en avant la situation d'une ou plusieurs années particulières. Ce que l'on commence à entrevoir sur cette vue mondiale, et que les vues régionale et nationale confirmeront, c'est essentiellement que la croissance de l'offre n'a pas suivi la croissance de la demande, c'est à dire que l'accès aux services d'eau potable et d'assainissement a cru moins vite globalement que la population mondiale, et en particulier que la population urbaine mondiale.

[SLIDE 13]

La seconde vue, régionale, se concentre sur des indicateurs agrégés pour chacune des 6 régions représentées dans nos données initiales, l'interactivité étant construite pour l'utilisateur au travers d'un filtre permettant de choisir la région et qui modifie tous les graphiques d'évolution historique, sauf ceux en bas à gauche et à droite qui ne représentent que l'année 2016 ; celui de gauche permet de comparer visuellement rapidement toutes les régions entre-elles du point de vue de l'accès aux services d'assainissement et de la mortalité totale pour les 2 sexes. L'utilisateur peut également zoomer sur un groupe de pays particuliers sur le graphique en nuage de points en bas à droite de la vue en utilisant les sliders de l'analyse bivariée qui compare l'accès à l'eau potable de chaque pays de la région sélectionnée avec sa stabilité politique en 2016. Comme nous l'avions pressenti en introduction de cette troisième partie, l'accès à l'eau potable est relativement dépendant de la stabilité politique du pays.

[SLIDE 14]

Cette analyse sera confirmée par notre troisième vue, la vue nationale, qui présente un certain nombre d'indicateurs permettant de situer le pays par rapport à la moyenne mondiale, particulièrement sur les gauges en bas de la vue pour l'accès à l'eau potable et l'accès à l'assainissement. On a également choisi de représenter pour l'année 2016 la population totale, le taux d'urbanisation, la mortalité par défaut d'assainissement ainsi que la population totale et la stabilité politique. On a aussi choisi de représenter

l’évolution de ces variables dans le temps sur le graphique du milieu – il faut préciser que toutes ces données n’existent pas pour tous les pays et qu’il n’est donc pas toujours possible de les représenter. Par exemple, si on choisit l’Europe puis la France, pour laquelle toutes les données existent, tous nos graphiques s’affichent, mais si on choisit par exemple l’Afrique et l’Angola, pays pour lequel certaines données n’existent pas, tous nos graphiques ne s’affichent pas nécessairement. Dans cette 3ème vue, l’interactivité est construite au travers de filtres qui permettent à l’utilisateur de choisir la région puis le pays à étudier.

[SLIDE 15]

En conclusion, il s’agira donc, pour le choix des pays pertinents, de trouver un équilibre entre la présence d’une stabilité politique suffisante pour permettre le succès du pilotage de projets d’infrastructure de long terme, tout en choisissant des pays où le taux d’urbanisation est en croissance significative et où les services d’eau potable et d’assainissement présentent des opportunités de développement. Pour mener à bien ce processus de décision, il conviendra d’enrichir les données de notre base, non seulement en les rafraîchissant pour les années postérieures à 2017, mais aussi en comblant l’information manquante et en ajoutant des données de nature économique et financière (croissance PIB par habitant et autres indicateurs de développement par exemple).